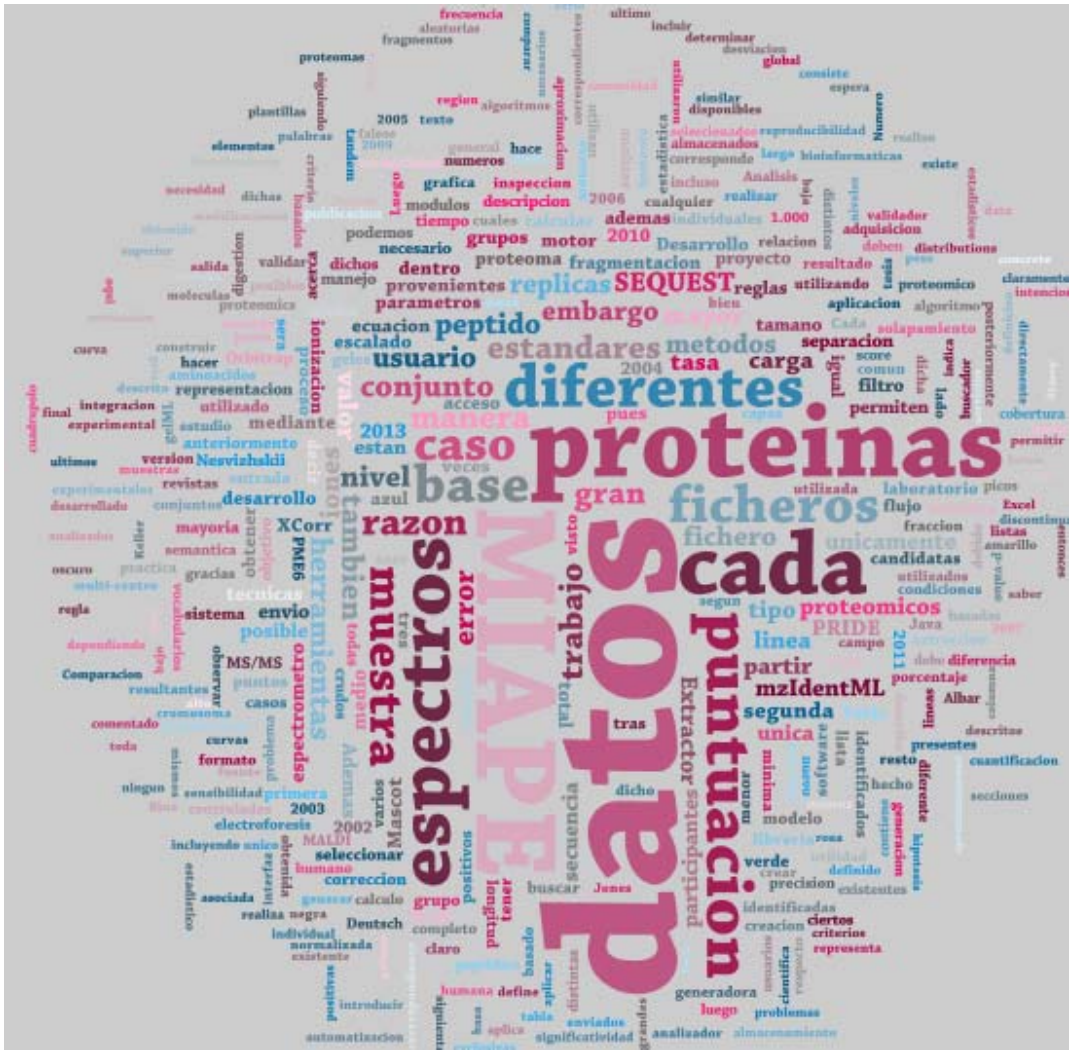
Métodos de validación de identificaciones a gran escala de proteínas y desarrollo e implementación de estándares en proteómica

Salvador Martínez de Bartolomé Izquierdo



Dedicado a Ainara, mi compañera de vida

Agradecimientos

Son ya más de diez años desde que entré en un laboratorio de Proteómica. Han pasado muchas cosas desde entonces, la gran mayoría muy buenas, y es ahora tiempo de recordar y plasmar aquí unas cuantas de ellas, mostrando también mi agradecimiento a tanta gente que me ha acompañado durante todo este tiempo. Lo cierto es que me siento muy afortunado por haber recorrido este camino y por haber conocido a tanta buena gente en él.

Recuerdo perfectamente el día de la entrevista con Jesús en la que me hizo un croquis en papel de la proteómica en general. ¿Quién iba a saber que ese iba a ser el principio de todo? Agradezco en primer lugar a Jesús la confianza que enseguida depositó en mí y lo mucho que disfruté mi trabajo con él durante casi tres años, gracias a esa pasión tan característica y pegadiza con la que se enfrentaba a los problemas y enigmas de la Proteómica todos los días.

Luego, pasé a formar parte del laboratorio de Juan Pablo en el que he trabajado ya casi ocho años. Siempre le agradeceré la apuesta que hizo por la bioinformática para el proyecto, junto con Jose María Carazo al principio, siendo el mejor embajador que uno puede tener de mi trabajo por todo el mundo. Además, uno de los aspectos que más valoro de estos años es el haber podido viajar y conocer a tanta gente tanto en España como en el extranjero. Gracias Juan Pablo por tu apoyo en todo este tiempo, por darme todas las facilidades que me has dado y por saber valorar el trabajo realizado. Gracias también por apoyarme y animarme para el nuevo camino se acerca.

Tanto en un laboratorio como en otro siempre me he encontrado a muy buenos compañeros y amigos, con los que he compartido durante todo este tiempo alegrías, tristezas, frustraciones, cabreos con el ordenador, risas y fiestas. Dani, Alberto Jorge, Anabel, Yoli, Espe, Merche, y Marga qué bien lo pasábamos gracias a vuestro buen humor. Recuerdo Marga cuando me contabas tipo abuela cebolleta que tu tesis fueron 10 años…y yo pensaba madre mía… pues mírame ahora… Luego en el CNB Alberto y Miguel Ángel, mis compañeros de batalla. Gracias Alberto por haber sido, sobre todo al principio, el modelo de muchas cosas que he aprendido en lo profesional y personal y gracias Miguel Ángel por estar siempre tan dispuesto a ayudar y aunque quizás no seas consciente, has sido un gran apoyo en muchos momentos. Y gracias Emilio porque aunque estuvimos poco tiempo juntos, me enseñaste mucho acerca del maravilloso mundo del Java y Eclipse. Compañeros del laboratorio como Silvia, pocas veces me ha dado más pena que alguien se fuese, Virginia, con sus "paseos", gracias por preocuparte siempre por mis cosas, y Lucia, la persona que más en el centro de Madrid ha vivido nunca, completando el grupo de "tinieblas en pelotas". Gran momento ese ;). Rosana, que ya desde mi etapa en el CBM nos conocimos y siempre me gustó mucho tu forma de ser. Antonio, igualmente conocido desde el principio, también cruzaste el charco de JV a JP, gracias por

responder a mis preguntas tan apasionantes sobre FDRs. Alberto Paradela…no puedo más que recordar ese momento tú y yo a metro y medio del suelo… Marisol, siempre con su sonrisa…gracias por introducir el término "luncherita" en mi vida. Miguel Marcilla, gracias por estar siempre también dispuesto a ayudar y responder cualquier pregunta, Adán, gracias por ofrecerme tu amistad y la de la triple "A", Carmen, principal usuaria del gran Extractor y compañera de mi última etapa en el B1, Severine, grandísima persona, Sergio, el de "Hueva" (como parece ser que me llamas tú) y sus consejos sobre cualquier compra de cualquier artículo que te imagines, Lola, la gerente más hippie que he tenido ;), Inés y sus "hola!" cuando me siento en el ordenador de al lado, Estefanía…o ya la he nombrado…ah no, gracias por ser tan natural…"esgueva"… y tantos otros compañeros y compañeras con los que he compartido mil momentos muy muy buenos.

También gracias a toda la gente que he conocido por toda España y con la que también he compartido un montón de momentos, de charlas, de preguntas, de risas, días en La Cristalera, mareos en autobús, partidas de ping-pong incluso con Peter Roepstorff y de futbolín (frota frota…), congresos miles, posters, fiestas,… Gracias también a gente de fuera como los compañeros del HUPO-PSI: Andy Jones, Juan Antonio, Pierre-Alain, Martin Eisenacher y Eric Deutsch, con los que ha sido un orgullo trabajar y viajar en busca de los estándares perdidos.

Fuera ya de la Proteómica quiero agradecer con toda mi alma a Ainara, a quien dedico todo este esfuerzo, por estar a mi lado siempre, por animarme siempre, por ser mi fan número uno siempre y por hacer que mis días sean mejores siempre.

A mi madre y a mi padre, por darme las oportunidades que tanto he aprovechado. Os quiero. A mi suegra Matilde, la mejor suegra que uno se puede imaginar. Y a grandes amigas como Sole y Mari Cruz, porque sé que siempre estáis ahí para lo que necesite.

También gracias a mis compis darderos, Nuria, Borja, Richy, Gema y Sole, por hacer que mi mente se despejase de cualquier agobio una vez por semana a costa, eso sí, de sueño y del bolsillo.

Y pese a que seguramente me deje a mucha gente sin nombrar aquí, quiero dar las gracias a todos porque gracias a vosotros he llegado donde estoy ahora y a ser lo que soy.

GRACIAS.

Facultad de Ciencias
Departamento de Biología Molecular

MÉTODOS DE VALIDACIÓN DE IDENTIFICACIONES A GRAN ESCALA DE PROTEÍNAS Y DESARROLLO E IMPLEMENTACIÓN DE ESTÁNDARES EN PROTEÓMICA

Memoria presentada para optar al grado de Doctor en Ciencias por el licenciado

Salvador Martínez de Bartolomé Izquierdo

Septiembre 2013

Directores de Tesis:

Juan Pablo Albar
Jesús Vázquez Cobos

Summary

High throughput identification of peptides in databases from tandem mass spectrometry data is a key technique in modern proteomics. Common approaches to interpret large scale peptide identification results are based on the statistical analysis of average score distributions, which are constructed from the set of best scores produced by large collections of MS/MS spectra by using searching engines such as SEQUEST. Other approaches calculate individual peptide identification probabilities on the basis of theoretical models or from single-spectrum score distributions constructed by the set of scores produced by each MS/MS spectrum. In this work, we study the mathematical properties of average SEQUEST score distributions by introducing the concept of spectrum quality and expressing these average distributions as compositions of single-spectrum distributions. Our analysis leads to a novel indicator, the probability ratio, a non-parametric and robust indicator that makes spectra classification according to parameters such as charge state unnecessary and allows a peptide identification performance, on the basis of false discovery rates, that is better than that obtained by other empirical statistical approaches. We also developed another method based on the construction of single-spectrum SEQUEST score distributions. These results make the robustness, conceptual simplicity, and ease of automation of the probability ratio algorithm a very attractive alternative to determine peptide identification confidences and error rates in high throughput experiments.

On the other hand, recent developments of HUPO-PSI (Proteomics Standards Initiative) standard data formats and MIAPE guidelines (Minimum Information About a Proteomics Experiment) are certainly contributing to proteomics data-sharing within the scientific community. In addition, specialized journals have emphasized the use of these standards and guidelines to facilitate the evaluation and publication of new articles. However, there is an evident lack of bioinformatics tools specifically designed to manage these standards containing the required information and its connectivity with the proteomics pipeline. In this work we describe the development of a set tools based on PSI standards and MIAPE guidelines, such as semantic and MIAPE validators of proteomics standard data files, a proteomics experiment repository based on MIAPE guidelines, a Java library for the management and extraction of MIAPE information from standard data files and a tool for a complete proteomics data analysis workflow allowing the aggregation, filtering and inspection of large amount of data, as well as its dissemination by preparing a complete ProteomeXchange submission. Additionally, here we also present the contribution for the definition of the MIAPE guidelines for quantitative

Proteomics experiments, receptly accepted as a new global standard for the Proteomics community.

Abreviaturas

1D	Una dimensión
2D	Dos dimensiones
2-DE	Electroforesis bidimensional
APCI	Ionización química a presión atmosférica
CID	Disociación inducida por colisión
CV	Vocabulario controlado
Da	Dalton
DIGE	Electroforesis diferencial en gel
ECD	Disociación por captura de electrones
EI	Ionización de impacto de iones
ESI	Ionización por electrospray
ETD	Disociación por transferencia de electrones
FAB	Bombardeo de átomos rápido
FDR	Tasa de falsos positivos
FT-ICR	Analizador de resonancia ciclotrónica de iones mediante transformada de Fourier
HPP	Proyecto del Proteoma Humano
HUPO	Organización del proteoma humano
IEF	Isoelectroenfoque
IT	Trampa iónica
LIMS	Sistema de manejo de información de laboratorio
MALDI	Ionización mediante desorción por láser asistida por matriz
MGF	Formato genérico de Mascot
MIAPE	Información mínima acerca de un experimento proteómico
MIAPE GE	Información mínima acerca de un experimento proteómico basado en electroforesis en gel
MIAPE GI	Información mínima acerca del análisis de las imágenes de un experimento

	proteómico basado en geles
MIAPE MS	Información mínima acerca de un experimento proteómico por espectrometría de masas
MIAPE MSI	Información mínima acerca del análisis informático de los datos de un experimento proteómico por espectrometría de masas
MIAPE Quant	Información mínima acerca de un experimento proteómico cuantitativo
MRM	Monitorización de reacción múltiple
MS	Espectrometría de masas / Espectro de masas
MS/MS	Espectrometría de masas en tándem / Espectro de fragmentación
PEP	Probabilidad de error a posteriori
PFF	Huella de espectros de fragmentación de péptidos
PME	Experimento multi-centro de ProteoRed
PMF	Huella de masas peptídicas
PSI	Iniciativa de estandarización en proteómica
PSM	Asignación entre péptido y espectro
PTM	Modificación post-traduccional
Q-TOF	Analizador de masas mediante cuadrupolo y tiempo de vuelo
QIT	Trampa de iones cuadrupolar
RP-HPLC	Cromatografía liquida de alto rendimiento por fase reversa
SDS-PAGE	Electroforesis en gel poliacrilamida con dodecil-sulfato sódico
TOF	Tiempo de vuelo
XML	Lenguaje de marcas extensible
XSD	Definición del esquema XML

Publicaciones generadas durante el periodo de elaboración de esta tesis

1. Vizcaino, J. A., E. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Ríos, J. A. Dianes, Z. Sun, T. Farrah, N. Bandeira, P. A. Binz, I. Xenarios, M. Eisenacher, G. Mayer, L. Gatto, A. Campos, R. J. Chalkley, H. J. Kraus, J. P. Albar, S. Martinez-Bartolomé, R. Apweiler, G. Omenn, L. Martens, A. R. Jones and H. Hermjakob (2013). "ProteomeXchange: globally coordinated proteomics data submission and dissemination." **Nat Biotechnol** (Submitted).
2. Segura, V., A. Medina-Aunon, M. Mora, S. Martínez-Bartolomé, J. Abian, K. Aloria, O. Antúnez, J. Arizmendi, M. Azkargorta, S. Barceló, J. Beaskoetxea, J. Bech-Serra, F. J. Blanco, M. Braga-Monteiro, D. Cáceres, F. Canals, M. Carrascal, J. I. Casal, F. Clemente, N. Colome, N. Dasilva, P. Díaz, F. Elortza, P. Fernández-Puente, M. Fuentes, O. Gallardo, S. Gharbi, C. Gil, M. Hernáez, M. Lombardia, M. Lopez-Lucendo, M. Marcilla, J. Mato, M. Mendes, E. Oliveira, I. Orera, A. Pascual, G. Prieto, C. Ruiz-Romero, M. Sánchez del Pino, D. Tabas-Madrid, M. Valero, V. Vialas, J. Villanueva, J. P. Albar and F. Corrales (2013). "Surfing transcriptomic landscapes. A step beyond the annotation of Chromosome 16 proteome." **Journal of Proteome Research** (Submitted).
3. Ghali, F., R. Krishna, P. Lukasse, S. Martinez-Bartolome, F. Reisinger, H. Hermjakob, J. A. Vizcaino and A. R. Jones (2013). "A toolkit for the mzIdentML standard: the ProteoIDViewer, the mzidLibrary and the mzidValidator." **Mol Cell Proteomics.**
4. Martinez-Bartolome, S., E. W. Deutsch, P. A. Binz, A. R. Jones, M. Eisenacher, G. Mayer, A. Campos, F. Canals, J. J. Bech-Serra, M. Carrascal, M. Gay, A. Paradela, R. Navajas, M. Marcilla, M. L. Hernaez, M. D. Gutierrez-Blazquez, L. F. Velarde, K. Aloria, J. Beaskoetxea, J. A. Medina-Aunon and J. P. Albar (2013). "Guidelines for reporting quantitative mass spectrometry based experiments in proteomics." **J Proteomics**.
5. Segura, V., J. A. Medina-Aunon, E. Guruceaga, S. I. Gharbi, C. Gonzalez-Tejedo, M. M. Sanchez del Pino, F. Canals, M. Fuentes, J. I. Casal, S. Martinez-Bartolome, F. Elortza, J. M. Mato, J. M. Arizmendi, J. Abian, E. Oliveira, C. Gil, F. Vivanco, F. Blanco, J. P. Albar and F. J. Corrales

(2013). "Spanish human proteome project: dissection of chromosome 16." **J Proteome Res** 12(1): 112-122.

6. Martinez-Bartolome, S., P. A. Binz and J. P. Albar (2013). The Minimal Information About a Proteomics Experiment (MIAPE) from the Proteomics Standards Initiative. **Plant Proteomics: Methods and Protocols, Second Edition.** J. V. Jorrin-Novo. New York, USA, Humana Press. 1072.
7. Orchard, S., J. P. Albar, E. W. Deutsch, M. Eisenacher, P. A. Binz, S. Martinez-Bartolome, J. A. Vizcaino and H. Hermjakob (2012). "From proteomics data representation to public data flow: a report on the HUPO-PSI workshop September 2011, Geneva, Switzerland." **Proteomics** 12(3): 351-355.
8. Medina-Aunon, J. A., S. Martinez-Bartolome, M. A. Lopez-Garcia, E. Salazar, R. Navajas, A. R. Jones, A. Paradela and J. P. Albar (2011). "The ProteoRed MIAPE web toolkit: a user-friendly framework to connect and share proteomics standards." **Mol Cell Proteomics** 10(10): M111 008334.
9. Kenyani, J., J. A. Medina-Aunon, S. Martinez-Bartolome, J. P. Albar, J. M. Wastling and A. R. Jones (2011). "A DIGE study on the effects of salbutamol on the rat muscle proteome - an exemplar of best practice for data sharing in proteomics." **BMC Res Notes** 4: 86.
10. Gibson, F., C. Hoogland, S. Martinez-Bartolome, J. A. Medina-Aunon, J. P. Albar, G. Babnigg, A. Wipat, H. Hermjakob, J. S. Almeida, R. Stanislaus, N. W. Paton and A. R. Jones (2010). "The gel electrophoresis markup language (GelML) from the Proteomics Standards Initiative." **Proteomics** 10(17): 3073-3081.
11. Hoogland, C., M. O'Gorman, P. Bogard, F. Gibson, M. Berth, S. J. Cockell, A. Ekefjard, O. Forsstrom-Olsson, A. Kapferer, M. Nilsson, S. Martinez-Bartolome, J. P. Albar, S. Echevarria-Zomeno, M. Martinez-Gomariz, J. Joets, P. A. Binz, C. F. Taylor, A. Dowsey, A. R. Jones and E. Minimum Information About a Proteomics (2010). "Guidelines for reporting the use of gel image informatics in proteomics." **Nat Biotechnol** 28(7): 655-656.
12. Martinez-Bartolome, S., F. Blanco and J. P. Albar (2010). "Relevance of proteomics standards for the ProteoRed Spanish organization." **J Proteomics** 73(6): 1061-1066.
13. Martinez-Bartolome, S., J. A. Medina-Aunon, A. R. Jones and J. P. Albar (2010). "Semi-automatic tool to describe, store and compare proteomics experiments based on MIAPE compliant reports." **Proteomics** 10(6): 1256-1260.
14. Martinez-Bartolome, S. and J. P. Albar (2009). "Estado actual de las directrices MIAPE para la

estandarización de informes relativos a experimentos proteómicos." **Proteómica** IV: 29-35.

15. Martinez-Bartolome, S., P. Navarro, F. Martin-Maroto, D. Lopez-Ferrer, A. Ramos-Fernandez, M. Villar, J. P. Garcia-Ruiz and J. Vazquez (2008). "Properties of average score distributions of SEQUEST: the probability ratio method." **Mol Cell Proteomics** 7(6): 1135-1145.
16. Lopez-Ferrer, D., A. Ramos-Fernandez, S. Martinez-Bartolome, P. Garcia-Ruiz and J. Vazquez (2006). "Quantitative proteomics using 16O/18O labeling and linear ion trap mass spectrometry." **Proteomics** 6 Suppl 1: S4-11.
17. Lopez-Ferrer, D., S. Martinez-Bartolome, M. Villar, M. Campillos, F. Martin-Maroto and J. Vazquez (2004). "Statistical model for large-scale peptide identification in databases from tandem mass spectra using SEQUEST." **Anal Chem** 76(23): 6853-6860.

Índice

Índice

Introducción

Introducción

A lo largo del desarrollo de esta tesis se describirán primeramente los algoritmos y aproximaciones desarrolladas en el laboratorio de Proteómica del Centro de Biología Molecular Severo Ochoa, CSIC-UAM entre los años 2003 y 2005, bajo la dirección de Jesús Vázquez, para la validación de identificaciones de péptidos y proteínas a gran escala. Posteriormente, se describirán las herramientas desarrolladas basadas en estándares de Proteómica correspondientes al periodo comprendido entre el 2006 al 2013 en el laboratorio de Proteómica del Centro Nacional de Biotecnología, CSIC, bajo la dirección de Juan Pablo Albar.

1.1. ¿Qué es la Proteómica?

Las proteínas son moléculas formadas por cadenas lineales de aminoácidos que representan la unidad funcional de las células siendo imprescindibles para una gran cantidad de funciones como pueden ser: estructurales, enzimáticas, homeostáticas, transducción de señales o inmunológicas.

El término "**proteoma**", acuñado por Marc Wilkins (Wasinger, Cordwell et al. 1995, Wilkins, Pasquali et al. 1996) se define como el conjunto completo de proteínas que podrían estar presentes en una muestra u organismo, derivadas de la transcripción de un genoma, o como el conjunto de proteínas que pueden ser detectadas por una determinada metodología experimental. La descripción del proteoma permite tener una imagen dinámica de todas las proteínas expresadas en un momento dado y bajo determinadas condiciones concretas de tiempo y ambiente. El estudio y comparación sistemáticos del proteoma en diferentes situaciones metabólicas y/o patológicas permite identificar aquellas proteínas cuya presencia, ausencia o alteración se correlaciona con determinados estadíos fisiológicos. En el caso concreto del análisis proteómico asociado a patologías concretas, es posible identificar proteínas que permitirían diagnosticar la enfermedad o pronosticar la evolución de la misma. Dichas proteínas son biomarcadores de dicha enfermedad.

El término "**Proteómica**" se refiere precisamente al estudio del proteoma, esto es, a un conjunto de proteínas presentes en una determinada muestra biológica bajo unas determinadas condiciones. Su objetivo es por tanto caracterizar o describir las proteínas ya sea de manera cualitativa o cuantitativa. Dentro del campo de la Proteómica se pueden distinguir diferentes áreas: la Proteómica descriptiva o cualitativa, que trata de identificar las proteínas que están presentes en una muestra, así como la caracterización de modificaciones post-traduccionales; la

Proteómica cuantitativa o "de expresión diferencial", que trata de describir los cambios en los niveles de expresión global de proteínas entre dos o más muestras; y el análisis de las interacciones entre proteínas, analizando las posibles interacciones entre diferentes proteínas y construyendo redes de interacciones como por ejemplo las cascadas de señalización.

Durante los últimos 30 años se han ido desarrollando técnicas complementarias que han favorecido los estudios proteómicos como son la electroforesis en gel, la cromatografía líquida o la espectrometría de masas. Sin embargo, ha sido en los últimos de 15 a 20 años cuando se han desarrollado aproximaciones que permiten la detección y/o cuantificación de muchas proteínas simultáneamente, acercándose más a métodos globales de análisis proteómicos. Este cambio de paradigma se debe en parte a los avances tecnológicos en el diseño de espectrómetros de masas e incluso más significativamente es debido a la disponibilidad de datos de secuencias genómicas de diferentes organismos con los cuales la espectrometría de masas es capaz de identificar péptidos y proteínas a gran escala. En el caso de la espectrometría de masas, destaca el desarrollo fundamental de la tecnología de ionización por electrospray (ESI, *Electrospray Ionization*) desarrollada por John B. Fenn (Fenn, Mann et al. 1989), ganador del premio Nobel en Proteómica en el año 2002, junto con Koichi Tanaka por su contribución con la tecnología de desorción por láser asistida por matriz (MALDI, *Matrix-Assisted Laser Desorption/Ionization)* (Karas, Bachmann et al. 1985), dos ionizaciones suaves que permitirían el estudio de las proteínas mediante la espectrometría de masas. Actualmente, los métodos de identificación permiten la detección de varios miles de proteínas con un coste económico y de tiempo relativamente bajo.

1.2. El espectrómetro de masas

El espectrómetro de masas es un instrumento que permite analizar con gran precisión la composición de diferentes sustancias químicas o bioquímicas, separando las moléculas en función de su relación masa-carga (m/z). El espectrómetro de masas no es un instrumento de reciente creación. Su origen se remonta a la primera mitad del siglo XX cuando Joseph John Thomson, descubridor del electrón en 1897 y galardonado por ello con el premio Nobel en Física (1906), construyó un instrumento similar a un espectrómetro de masas con la intención de medir la relación masa-carga del electrón (Thomson 1910, Thomson 1913). Se denominó espectrógrafo parabólico ya que era capaz de separar los iones por su trayectoria parabólica diferencial al ser sometidos a un campo electromagnético. Después en 1918, Francis W. Aston construyó un espectrómetro de masas con mayor resolución y que permitía el estudio de isótopos (Aston 1918, Aston 1919, Aston 1920). En el mismo periodo, Arthur Jeffrey Dempster aumentó la resolución mediante un analizador magnético y desarrolló la primera fuente de

impacto electrónico, que permitía la ionización de moléculas volátiles al ser expuestas a un haz de electrones (Dempster 1917).

La principal limitación existente por aquel entonces era tener la suficiente precisión para permitir el análisis de elementos y pequeñas moléculas.

De manera simplificada, la Figura 1 muestra los principales componentes presentes en un espectrómetro de masas:

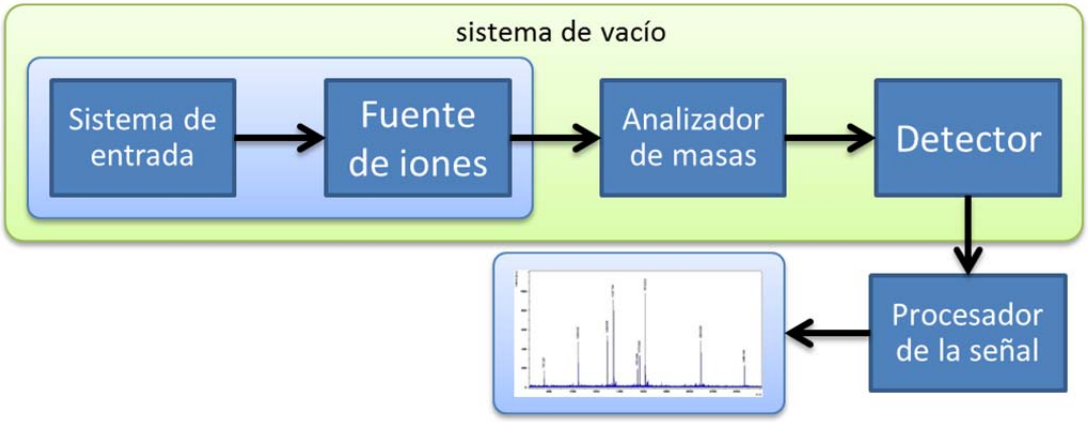


Figura 1. Componentes principales de un espectrómetro de masas: Sistema de entrada, fuente de iones, analizador de masas, detector y procesador de la señal.

A. **Sistema de entrada**: es el sistema para introducir una pequeña cantidad de muestra en el espectrómetro. A menudo este sistema contiene un medio que permite la volatilización de muestras sólidas o líquidas.
B. **Fuente de iones**: la fuente de ionización convierte los componentes de la muestra en iones (partículas cargadas) en fase gaseosa. En muchos casos, el sistema de entrada y la fuente de ionización están combinados en un único componente. En Proteómica las fuentes de ionización ESI y MALDI son las más utilizadas.

   Gracias a la introducción de técnicas de ionización suave, la espectrometría de masas se introdujo en el ámbito del análisis de biomoléculas al final de la década de los años 70 del siglo pasado. Algunas de esas técnicas son:

   - **Bombardeo de átomos rápido** (FAB, *Fast Atom Bombardment*) (Roepstorff y Richter 1992): el bombardeo se realiza con un haz de alta energía de átomos neutros (típicamente Xenón o Argón) que colisiona con la muestra causando la desorción y la ionización. Esta técnica de ionización se utiliza principalmente para el análisis de moléculas biológicas grandes las cuales son difíciles de pasar a fase gaseosa.

- **Desorción por campo eléctrico** (FD, *Field Desorption*)**:** Los analitos pierden un electrón cuando se les aplica un campo eléctrico lo suficientemente intenso, creado por un emisor de campo.

- **Desorción por plasma** (PD, *Plasma Desorption*) (Macfarlane 1990): La desintegración del Californio 252 (^{252}Cf) produce una partícula alfa y dos fragmentos cargados, típicamente ^{106}Te y ^{142}Ba. Estos dos fragmentos viajan en direcciones opuestas. Uno de esos fragmentos choca contra la muestra creando de 1 a 10 iones. El otro fragmento choca con el detector marcando el inicio de la adquisición de datos. Esta técnica es también utilizada para el análisis de moléculas biológicas grandes.

Estas técnicas permitían la ionización de moléculas termolábiles de gran tamaño, como los péptidos y las proteínas, sin producir una extensiva degradación característica de las denominadas técnicas de ionización dura, cuyo representante de referencia es el impacto electrónico de la ionización química:

- **Ionización química**: la ionización química utiliza iones reactivos para interaccionar con los analitos. Esto se consigue mediante la introducción de un exceso de metano en la fuente de ionización de impacto de iones (EI, *Electron Impact*).

En especial el FAB (Barber, Bordoli et al. 1981) constituyó la herramienta básica para el análisis y secuenciación de péptidos durante más de una década. Posteriormente, en el inicio de los años 90, otros dos métodos de ionización suave, el electrospray (**ESI,** *Electrospray Ionization*) y la desorción por láser asistida por matriz (**MALDI,** *Matrix-Assisted Laser Desorption/Ionization*), tomaron el relevo al FAB y constituyeron los pilares básicos de la espectrometría de masas de la proteómica contemporánea.

- **Ionización (MALDI)**: La muestra se implanta en una matriz sólida, a menudo consistente en material orgánico, la cual protege a la muestra de ser destruida, además de facilitar su evaporación e ionización. Entonces es radiada con un láser pulsado (p.ej. un láser de nitrógeno), lo cual hace que la matriz expulse iones, cationes y macromoléculas neutras, que crean una nube de gas que ionizará las macromoléculas de la muestra. Normalmente, esta técnica de ionización es utilizada junto con un analizador de tiempo de vuelo (TOF, *Time of Flight*), ya que al ser una técnica de ionización por pulsos, permite saber el tiempo en el que el compuesto a analizar se ioniza y empieza su viaje hacia el detector (Karas, Bachmann et al. 1985).

- **Ionización por electrospray (ESI)**: Se introduce el analito (que será ionizado) disuelto en un solvente más volátil por un capilar de metal muy pequeño y cargado. Debido a la repulsión de las cargas eléctricas, el líquido sale del capilar y forma un aerosol, una nube de pequeñas gotas (10 μm) altamente cargadas. Conforme el solvente se evapora, las moléculas de analito se aproximan, se repelen y finalmente, cuando la repulsión de las cargas positivas vence la tensión superficial, estallan las gotas (explosión de Coulomb). El proceso se repite hasta que el analito está libre de solvente, de modo que no quedan más que iones en fase gaseosa.

C. **Analizador de masas**: es la parte del espectrómetro que se encarga de separar los iones en función de su relación masa-carga (m/z). El analizador determina la resolución del espectrómetro, siendo ésta la capacidad del espectrómetro para distinguir entre masas próximas. El incremento continuo de las prestaciones de los analizadores de iones ha ido marcando y ampliando los límites con los que el investigador ha podido abordar el estudio del proteoma mediante espectrometría de masas en los últimos años.

- **Analizador por sector magnético** (Johnson y Nier 1953): este analizador acelera los iones provenientes de la fuente de ionización a una gran velocidad. Los iones pasan después por un sector magnético por el cual un campo magnético se aplica perpendicularmente a la trayectoria del ion. Esto produce una trayectoria circular del ion, la cual dependerá de la fuerza del campo magnético aplicado. Aplicando la ley de la fuerza de Lorentz, la relación masa-carga es calculada en función de la fuerza del campo magnético, del radio descrito por el ion y de su energía cinética.

- El **cuadrupolo** (Paul y Steinwedel 1953): El cuadrupolo consiste en cuatro barras paralelas de sección hiperbólica en la cara interna, generalmente de unos 15-20 cm de largo y 0.5 cm de radio, separadas entre sí unos 2 cm, a las que se aplica un potencial combinado de corriente continua y de radiofrecuencia que crean en su interior un campo denominado cuadrupolar que permite el paso únicamente de ciertas relaciones masa-carga de interés, actuando como un filtro, de forma que para una combinación de potenciales sólo los iones en un estrecho rango de valores m/z llega a incidir en el detector. En consecuencia, una pequeña fracción del total de iones es monitorizada mientras que el resto se desecha. Este modo de funcionamiento repercute negativamente en el límite de detección de estos instrumentos, especialmente cuando se requiere la obtención de espectros completos en rangos de masa amplios. Cuando a un cuadrupolo se le aplica únicamente el

potencial de radiofrecuencia, este sistema actúa como un filtro de banda ancha mucho más eficiente que los formados por lentes electrostáticas en analizadores de baja energía. Este modo de trabajo, conocido como *RF-only*, es el que utilizan las cámaras de colisión en los instrumentos de triple cuadrupolo y en otros espectrómetros en tándem que emplean colisiones de baja energía. Además, en los sistemas de **triple cuadrupolo**, cuando se realizan barridos convencionales (MS) uno de los dos analizadores debe trabajar también en modo RF-only.

- **Analizador por tiempo de vuelo** (TOF, *Time-Of-Flight*) (Stephens 1946): este analizador mide el tiempo transcurrido entre la salida de un ion de la fuente de ionización y su llegada al detector. La diferencia entre ambos tiempos es el llamado tiempo de vuelo y es proporcional a la relación masa-carga del ion, ya que los iones adquieren diferentes velocidades según ese valor y por tanto, tardan distinto tiempo en recorrer una determinada distancia. A diferencia de los sistemas de cuadrupolo o de sectores, que como se indicó anteriormente filtran en cada instante grupos de iones dentro de un pequeño rango de valores m/z desechando el resto de la población de iones, el analizador TOF separa y detecta en una escala de tiempo (tiempo de vuelo) el paquete completo de iones procedente de la fuente. El sistema trabaja por este motivo en régimen discontinuo por lo que es un detector indicado para técnicas de ionización de carácter pulsante como el MALDI.

- Las **trampas de iones** (IT, *Ion Trap*) o trampas de iones cuadrupolares (QIT, *Quadrupole Ion Trap*) permiten el confinamiento de los iones dentro de una cámara de pequeño tamaño utilizando campos eléctricos (*Paul trap* o *Ion trap*, *Orbitrap*) o magnéticos (analizador de resonancia ciclotrónica de iones mediante transformada de Fourier, FT-ICR, *Fourier Transform Ion Cyclotron Resonance*). Este tipo de analizadores permiten almacenar, seleccionar y analizar los iones formados en la misma trampa o en fuentes de ionización externas. Los iones pueden mantenerse en el interior de la trampa durante tiempos prolongados con objeto de favorecer la observación de descomposiciones metaestables o de fragmentos producidos por colisión con moléculas de gas. Un rasgo adicional de la trampa es que ésta permite aislar iones individuales que luego pueden, mediante la aplicación de un voltaje de radiofrecuencia, excitarse para su fragmentación con moléculas de Helio introducidas en la trampa. El primer instrumento basado en el QIT dirigido al análisis de biomoléculas (ESQUIRE) fue presentado por Bruker

Instruments en 1994. Este sistema podía utilizar diversos métodos de ionización como el ESI, el MALDI o el FAB. En 1995, Finnigan MAT introdujo LCQ, un sistema especialmente diseñado para ionización por ESI e ionización química a presión atmosférica (APCI). Para entonces el QIT había superado muchos de sus inconvenientes iniciales (rango de masas bajo, limitación al análisis de iones positivos) y mostraba las ventajas únicas que los sistemas de confinamiento de iones aportaban, entre ellas, su capacidad de acumulación de iones y de realizar análisis en tándem múltiple. La aparición de las fuentes de ESI-micro (Emmett y Caprioli 1994) ESI-nanospray (Wilm y Mann 1994) en 1995 hizo de estos sistemas uno de los analizadores más populares para el análisis de biomoléculas en las últimas dos décadas.

- La **trampa de iones lineal** (LIT, *Linear Ion Trap*) aísla los iones por medio de un sistema de cuadrupolo provisto en los extremos de sus barras de unos electrodos terminales con la misma geometría que éstas pero más cortos y aislados eléctricamente. A estos electrodos se les aplica un potencial que actúa como barrera para los iones atrapados en su interior. Este tipo de analizadores, denominados trampas iónicas lineales (LIT) o cuadrupolos de confinamiento lineal (LTQ, *Linear Trap Quadrupole*), mejoraron muchas de las características de los QITs aumentando su capacidad de almacenamiento de iones, velocidad de barrido y sensibilidad (Schwartz, Senko et al. 2002). La trampa LIT puede utilizarse además como un cuadrupolo convencional por lo que, situada como el segundo analizador en sistemas de triple cuadrupolo, permite trabajar en cualquiera de los modos de espectrometría de masas en tándem posibles en este tipo de instrumentos (barridos de precursores, pérdidas neutras y de iones producto) además de los modos propios de la trampa.

- El **Orbitrap** fue desarrollado y patentado por Makarov a finales de los años 90 (Makarov 2000) y consiste en una barra en forma de huso dentro de un cilindro cuyas paredes internas tienen esta misma forma (Hu, Noll et al. 2005). Los iones introducidos de forma perpendicular en este sistema adquieren un movimiento radial alrededor del huso combinado con un movimiento axial periódico cuya frecuencia es una función de su valor m/z. El Orbitrap es quizás el analizador de iones más joven en espectrometría de masas, aunque por sus características se ha introducido rápidamente en el área de la Proteómica. En un principio se comercializó acoplado a un

instrumento de trampa lineal que se utiliza para la acumulación de iones y su inyección en el Orbitrap, así como para llevar a cabo la fragmentación de iones en alguno de sus modos de espectrometría de masas en tándem. Actualmente también se comercializa acoplado a un cuadrupolo, lo que ofrece tiempos de ciclo mucho más rápidos y posibilidad de hacer multiplexing.

- **La resonancia iónica de ciclotrón con transformada de Fourier** (FT-ICR, *Fourier Transform Ion Cyclotron Resonance*) (Comisarow y Marshall 1974): se basa en la medición de la frecuencia de ciclotrón del movimiento circular descrito por los iones movidos por un campo magnético, la cual es dependiente de la relación masa-carga (Comisarow y Marshall 1974, Comisarow y Marshall 1996, Senko, Hendrickson et al. 1996).

D. **Detector**: el detector convierte el haz de iones en una señal eléctrica que puede ser procesada y almacenada. El detector más utilizado es el multiplicador de electrones el cual hace incidir el haz sobre un cátodo, liberándose electrones. Después de una serie de dínodos colocados a potenciales crecientes, se amplifica la corriente de electrones.

Como resultado del análisis de la muestra por el espectrómetro de masas obtenemos los espectros de masas, consistentes en una lista de masas, en realidad, de masas divididas por la carga del ion (m/z), asociadas con las intensidades con las que se detectaron. Se suelen representar por medio de un gráfico en el que en el eje de abscisas tenemos los valores m/z y en el eje de ordenadas las intensidades (Figura 2).

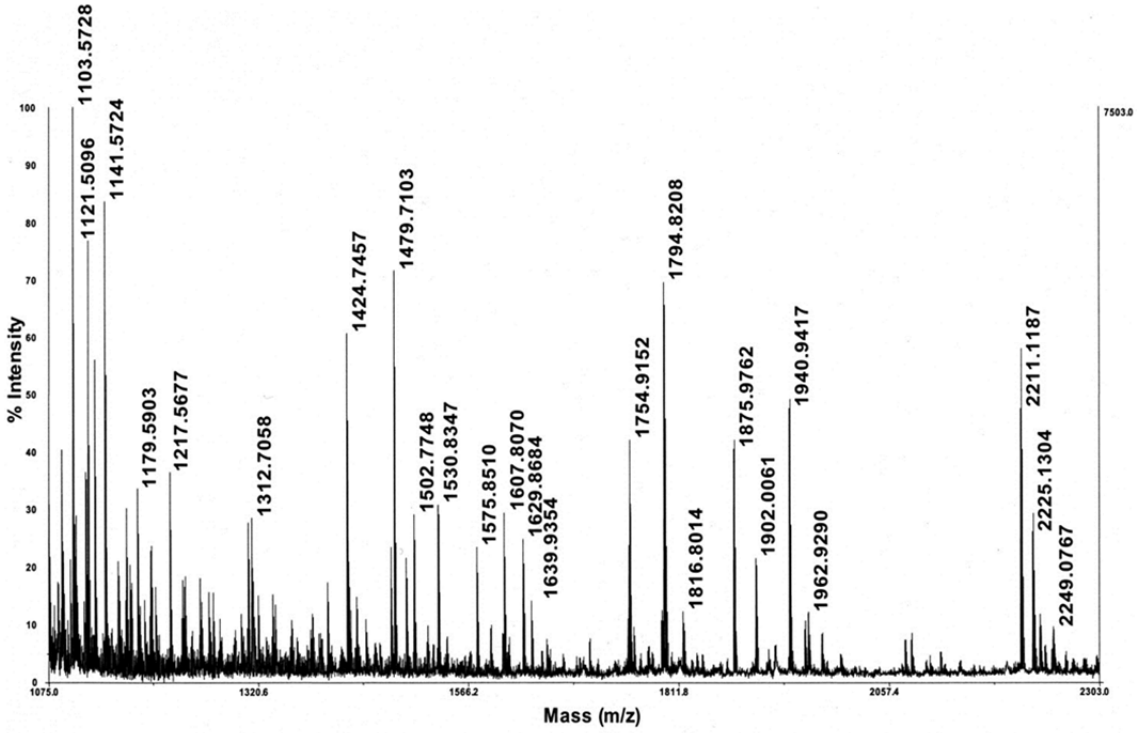


Figura 2. Espectro de masas (MS). En el eje de abscisas se muestra la relación masa/carga, en este caso, de los péptidos. En el de ordenadas, en este caso, el porcentaje de la intensidad de la señal de los picos.

1.3. La Proteómica clásica

Tradicionalmente, la Proteómica clásica o de primera generación se ha basado en tres técnicas principales: la separación y selección de las proteínas diana por medio de electroforesis en gel, la digestión de la proteína por medio de una endo-proteasa de especificidad de rotura conocida para obtener sus péptidos representativos, la adquisición de firmas moleculares, listas de masas moleculares o espectros por espectrometría de masas, y la identificación de proteínas por medio de búsquedas en bases de datos de los vínculos experimentales con éstas mediante herramientas bioinformáticas especializadas.

1.3.1. Electroforesis en gel

La electroforesis en gel es un conjunto de técnicas utilizadas para separar moléculas (en este caso proteínas), basándose en propiedades físico-químicas como su tamaño, su forma o su punto isoeléctrico, constituyendo así un proceso de purificación (Klose 1975, O'Farrell 1975).

El isoelectroenfoque (IEF, *isoelectric focusing*) (Bjellqvist, Ek et al. 1982, Gorg, Postel et al. 1988) permite separar las proteínas por su punto isoeléctrico por medio de un gradiente de pH y un campo eléctrico que se aplica a un gel, creando una diferencia de potencial entre los

dos extremos del gel. La electroforesis en geles con una matriz de poliacrilamida, comúnmente denominada electroforesis en poliacrilamida (PAGE, *polyacrilamide gel electrophoresis*) es una de las técnicas más comúnmente utilizadas para caracterizar mezclas complejas de proteínas. El gel constituye un entramado generado por la acrilamida y bis-acrilamida que dificulta el desplazamiento de las bio-moléculas. Las proteínas tienen la propiedad de desplazarse cuando se someten a un campo eléctrico ya que presentan una carga eléctrica neta si se encuentran en un medio que tenga un pH diferente al de su punto isoeléctrico. La velocidad de migración es inversamente proporcional a la masa. Las proteínas de mayor tamaño tendrán más dificultad en moverse debido al entramado de acrilamida. Cuando el pH del gel se iguala al de la proteína, ésta se detiene (se enfoca).

En una electroforesis nativa se somete a las proteínas a migración sin desnaturalización. En esta situación las proteínas migran en función de su carga, de su tamaño y de su forma. Además se mantienen en ciertos casos las interacciones entre subunidades y entre proteínas, separándose los complejos.

Por otro lado, en una electroforesis en condiciones desnaturalizantes, la más común, se somete a las proteínas a migración asegurando la completa desnaturalización (pérdida de la estructura tridimensional: estructuras secundarias y terciarias). En esta situación la migración es proporcional al tamaño de la molécula pero no a su forma. El agente desnaturalizante más empleado es el dodecil-sulfato sódico o SDS, un detergente (Laemmli 1970, Schagger y von Jagow 1987), que confiere a las proteínas una carga negativa en proporción a su masa. Las cargas negativas sobrepasan en número a las cargas positivas, impidiendo que éstas reviertan a la proteína a la conformación original. Las moléculas de SDS se unen a las proteínas de manera proporcional a su longitud o peso molecular por lo que las proteínas se cargan negativamente de forma proporcional a su masa, y por tanto migrarán en función de su peso molecular.

En una electroforesis mono-dimensional (1D), la separación se realiza mediante SDS-PAGE, y se puede obtener una separación de hasta 100 bandas. En una electroforesis bidimensional (2D) se realiza una primera dimensión con isoelectroenfoque y una segunda dimensión con SDS-PAGE y permite una separación de más de 2.000 manchas (*spots*) en un solo gel.

En cualquier tipo de electroforesis en gel, las moléculas deben ser posteriormente teñidas para poder visualizarlas. Para ello se realizan diferentes métodos de tinción, siendo los más utilizados las tinciones con plata (Gharahdaghi, Weinberg et al. 1999), con azul de coomassie (Schagger, Aquila et al. 1988) o con fluoróforos (*Sypro Ruby*). Una vez teñidas las bandas o las manchas de los geles, se pueden cortar físicamente para luego ser analizadas por el espectrómetro de masas.

1.3.2. Digestión

Previa a la identificación por espectrometría de masas, las proteínas se someten a un proceso de digestión por medio de una endo-proteasa (típicamente tripsina), mediante la cual la proteína es cortada por diferentes sitios en su secuencia (en el caso de la tripsina, siempre después de Lisina/Lys o Arginina/Arg). Este proceso de digestión resultará en un conjunto de péptidos con un rango de masas más adecuado para una posterior separación cromatográfica, y estos péptidos resultantes son analitos mucho menos complejos, adecuados para poder interpretar sus respectivos espectros de fragmentación. Normalmente, el patrón de corte de la proteasa es conocido, por lo que el conjunto de péptidos esperable a partir de un proteoma concreto puede predecirse a partir de las proteínas depositadas una base de datos.

Hay que tener en cuenta también que el proceso de digestión constituye en sí una fuente de error, debido a que la eficiencia de corte de las endo-proteasas nunca será del 100%, lo que supone un problema sobre todo para análisis cuantitativos posteriores. Además, idealmente los péptidos resultantes serán lo suficientemente específicos como para diferenciar unas proteínas de otras, sin embargo, dependiendo de la muestra, esto no siempre se cumple y en ocasiones una misma secuencia peptídica está presente en dos proteínas distintas. Esto último causa el problema de inferencia de proteínas, que será discutido más adelante (sección 1.4.4).

1.3.3. Espectrometría de masas (MS y MS/MS)

Una vez digerida la muestra proteica, ésta puede ser identificada en una sola etapa mediante la técnica de la huella de masas peptídicas (PMF, *Peptide Mass Fingerprinting*) (Henzel, Billeci et al. 1993, James, Quadroni et al. 1993, Mann, Hojrup et al. 1993, Pappin, Hojrup et al. 1993). Esta técnica se basa en la detección de las masas de los péptidos de la proteína problema, la cual ha tenido que ser aislada previamente, obteniéndose así una lista de masas de péptidos, la cual forma el patrón de masas o la huella peptídica asociada a dicha proteína (Figura 3). Esos patrones de masas son posteriormente comparados con los patrones teóricos de las masas de las secuencias peptídicas derivadas de la digestión teórica de las proteínas existentes en una base de datos (Aebersold y Goodlett 2001). La identificación de la proteína se produce con aquella existente en la base de datos que obtenga la mejor correlación de péptidos (masas peptídicas teóricas).

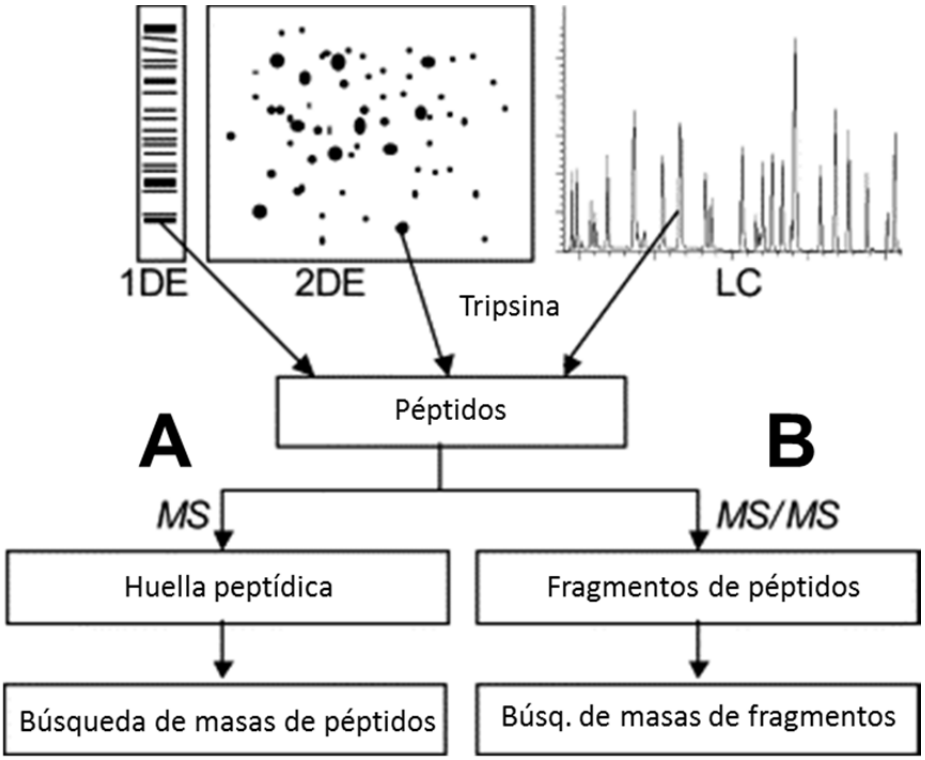


Figura 3. Flujos de trabajo en Proteómica: a partir de una banda de un gel mono-dimensional (1DE), un spot de un gel bidimensional (2DE) o una separación por cromatografía líquida (LC), se digieren los péptidos y son analizados, o bien por (A) huella peptídica (MS) en la que se realiza una búsqueda de las masas de los péptidos, o por (B) fragmentación (MS/MS o MS^2) en la que se realiza una búsqueda de las masas de los fragmentos de los péptidos.

Otra de las técnicas se basa en la identificación de los péptidos a partir de los espectros de fragmentación (PFF, *Peptide Fragment Fingerprinting*) (Mortz, O'Connor et al. 1996). En este caso, no es necesario el aislamiento de las proteínas a analizar. El análisis consta de dos fases: en la primera (MS), de la misma manera que con las huellas peptídicas se adquieren los espectros de masas de los péptidos (espectros MS); en una segunda fase (MS^2 o MS/MS), algunos de esos péptidos analizados serán fragmentados y las masas de los productos de dichas fragmentaciones serán analizadas de la misma manera comparándolas con las masas teóricas de las bases de datos (Figura 3). Es por estas dos fases por lo que se llama también espectrometría de masas en tándem. En este caso, el espectro de fragmentación contiene la información necesaria para secuenciar el péptido, pudiéndose obtener la secuencia de aminoácidos por medio del análisis de las diferencias entre las masas de los iones de fragmentación, es decir, de las masas de los picos principales del espectro (Figura 4).

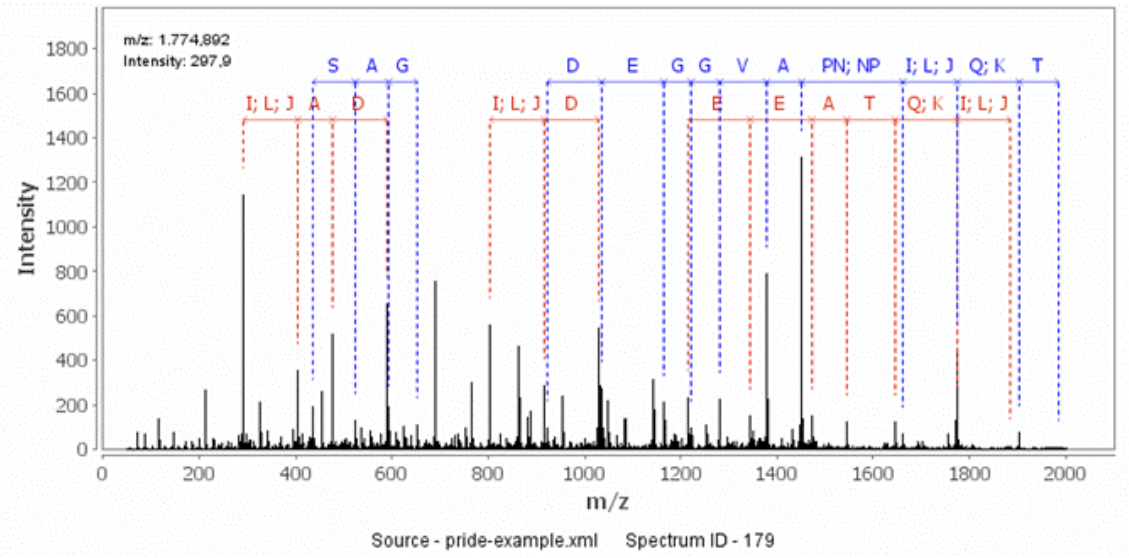


Figura 4. Espectro de fragmentación de masas (MS/MS). En el eje de abscisas se muestra la relación masa/carga, en este caso de fragmentos de péptidos. En el eje de ordenadas se muestra la intensidad absoluta de la señal de los picos.

La fragmentación de péptidos se realiza por diferentes metodologías tipo CID (*collision induced dissociation*), ETD (*electron transfer dissociation*), ECD (*electron-capture* dissociation), PSD (*post source decay*), entre otras. La disociación o fragmentación de iones, produce diferentes tipos de iones, los cuales se nombran siguiendo la notación definida por Roepstorff (Roepstorff y Fohlman 1984) y posteriormente modificada por Johnson (Johnson, Martin et al. 1987) con diferentes letras dependiendo del lugar de fragmentación y diferentes números dependiendo de la posición del aminoácido fragmentado (Figura 5).

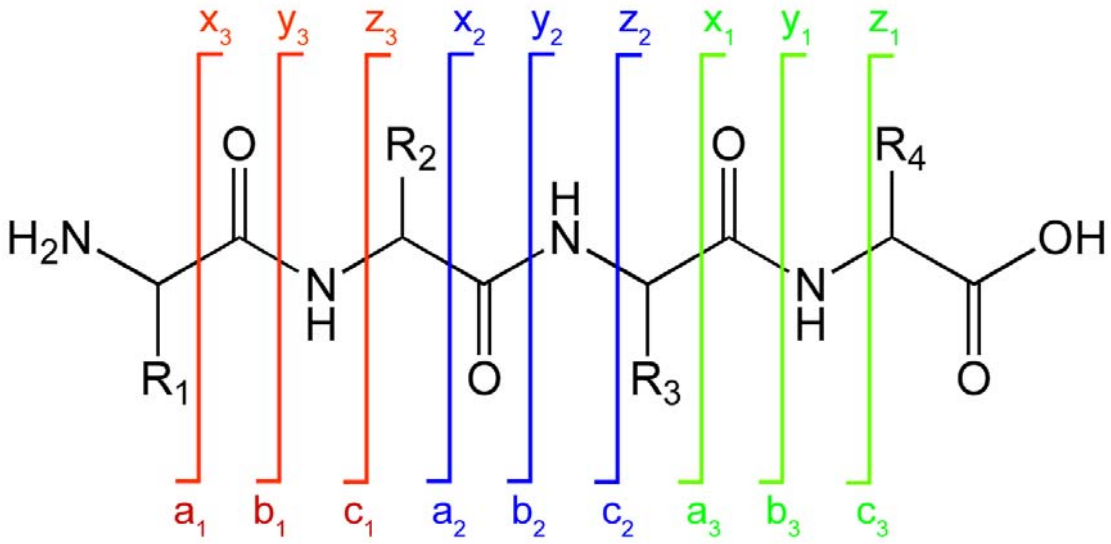


Figura 5. Nomenclatura según Roepstorff de la fragmentación de los péptidos. Dependiendo del número de residuos que tenga cada fragmento se diferencian diversas series, representadas con diferentes colores. Los números indican el número de residuos de aminoácidos contenidos en el ion al que se hace referencia, y la letra (a, b, c) o (x, y, z) indica el sitio de fragmentación del ion: las letras b e y corresponden a la ruptura en el enlace peptídico, la ruptura más habitual en activación CID.

Los fragmentos solo se podrán detectar en el espectrómetro si están cargados. Si el fragmento conserva el residuo C terminal del péptido, el ion será x, y o z. Si por el contrario conserva el N terminal, será de tipo a, b o c

1.4. Proteómica de segunda generación

Pese a la aparición de técnicas de separación en geles en dos dimensiones con las que se conseguía separar muestras complejas de proteínas, la Proteómica clásica seguía teniendo dificultades para detectar proteínas con un punto isoeléctrico extremo, de gran tamaño o inferior a 10 KDa o proteínas de membrana, así como ciertos problemas de reproducibilidad. Además presentaba dificultades para la detección de proteínas de baja concentración en la muestra en presencia de otras proteínas mucho más abundantes, teniendo un rango dinámico muy limitado (Gygi, Corthals et al. 2000). Sin embargo, la mayor limitación de esta técnica es que no es capaz de detectar más de un millar de proteínas en un solo gel, que en el caso del proteoma humano es de entorno al 10%. En el caso de la proteómica de segunda generación, se obtienen coberturas mucho mayores, superándose la cifra de 10.000 proteínas todas ellas identificadas y cuantificadas.

Para resolver esos problemas, la Proteómica de segunda generación o Proteómica moderna, llamada en inglés "*shotgun proteomics*", tiene por objetivo el análisis de muestras complejas

por medio de la combinación de una digestión en solución de los extractos proteicos con endo-proteasas, perdiéndose así la referencia a nivel de la proteína, seguida de una separación de alto rendimiento por cromatografía líquida (HPLC, *High Performance Liquid Cromatography*) de los péptidos generados, combinada con la espectrometría de masas MS/MS, llamada espectrometría de masas en tándem (*tandem mass spectrometry*) (Washburn, Wolters et al. 2001). Luego se utilizan los motores de búsqueda en bases de datos, herramientas bioinformáticas avanzadas que completan el proceso de identificación de proteínas a partir de los péptidos presentes en las muestras analizadas. La Proteómica de segunda generación permite realizar un análisis global y sistemático de los proteomas obteniéndose un mapa exhaustivo de las proteínas existentes en una muestra al poder utilizar el poder de separación de péptidos mediante técnicas cromatográficas acopladas directamente al espectrómetro de masas, permitiendo realizar medidas de miles de proteínas sin necesidad de realizar una separación de proteínas previa (*offline*) que, en la mayoría de los casos, es costosa en tiempo y material, además de poco reproducible. Esto posibilita análisis a nivel superior, como por ejemplo análisis de interacciones o de cascadas de señalización.

El estudio a gran escala de proteínas ha requerido de la evolución convergente de 1) técnicas de separación de proteínas y/o péptidos, 2) de técnicas de identificación masiva mediante espectrometría de masas, 3) de un crecimiento espectacular de las bases de datos generadas por los proyectos de secuenciación masiva de genomas, así como de 4) un desarrollo sin precedentes de algoritmos computacionales que vinculan los datos experimentales con los lenguajes propios de las bases de datos.

Actualmente la identificación de varios miles de proteínas es posible en un solo experimento (Wisniewski, Zougman et al. 2009, Beck, Schmidt et al. 2011, Nagaraj, Wisniewski et al. 2011, Nagaraj, Kulak et al. 2012, Picotti, Clement-Ziza et al. 2013), cuando hace poco más 15 años la simple identificación de una proteína era ya un éxito (Henzel, Billeci et al. 1993, James, Quadroni et al. 1993, Mann, Hojrup et al. 1993, Pappin, Hojrup et al. 1993), lo que evidencia la gran evolución que ha experimentado la Proteómica.

1.4.1. Separación multidimensional de péptidos

El rendimiento en la separación de los péptidos resultantes de la digestión determinará la calidad de las identificaciones debido a que se minimizará el solapamiento de señales de diferentes péptidos coeluyentes (péptidos que se separarían de forma similar y entrarían en el analizador de masas en el mismo rango de tiempo), lo cual reducirá el número de espectros quiméricos (formados por los fragmentos producto de más de un tipo de molécula). La complejidad de muestras proteicas provenientes de lisados celulares o fluidos biológicos

requiere de varias etapas de separación ortogonal (basadas en propiedades físico-químicas distintas).

Una de las primeras aproximaciones consiste en digerir mezclas de complejidad media de proteínas sin separación previa, y los péptidos resultantes se pre-fraccionan mediante distintas etapas de cromatografía con columnas de intercambio iónico (habitualmente catiónico), recolectando las fracciones que posteriormente se analizan mediante nano-cromatografía en fase reversa acoplada a un espectrómetro de masas en tándem (RP-HPLC-MS/MS) (Link, Eng et al. 1999). Otra estrategia similar es la separación de proteínas mediante SDS-PAGE mono-dimensional (fraccionamiento del proteoma por tamaños), seguida de la digestión por separado en diversas zonas del gel (hasta, por ejemplo 30 fracciones) y del análisis de cada fracción mediante nano-RP-HPLC-MS/MS (GelLCMS) (hidrofobicidad) (Lundby y Olsen 2011). Otras aproximaciones implican técnicas de separación de los péptidos por isoelectroenfoque (IEF) en solución, Off-Gel (Horth, Miller et al. 2006) (separación por carga), seguida de nano-RP-HPLC/UPLC-MS/MS habiendo demostrado ser altamente reproducibles y permitir un rendimiento óptimo (de Godoy, Olsen et al. 2008).

1.4.2. Aproximaciones computacionales para la identificación de proteínas a partir de espectros de fragmentación

Motores de búsqueda y tipos de puntuaciones

Para poder interpretar los espectros de fragmentación de masas y poder asignarlos a secuencias peptídicas, se utilizan diversos programas bioinformáticos llamados motores de búsqueda (*search engines*). Estos motores de búsqueda realizan una digestión teórica de las proteínas de una base de datos de secuencias proteicas, lo cual resulta en un conjunto de secuencias peptídicas, así como la fragmentación teórica de dichos péptidos. Cada una de dichas secuencias se asocia a un espectro de fragmentación teórico. Para cada espectro de fragmentación experimental, realizan una búsqueda de la secuencia peptídica que contenga el patrón de masas de fragmentación teórico más parecido al experimental. Cada motor de búsqueda realiza esta comparación de manera diferente, basándose en diferentes métricas. Por último, a cada asignación de una secuencia a un espectro, lo que llamaremos un PSM (*Peptide Spectrum Match*) se le asocia un valor de confianza, que puede ser una puntuación o una probabilidad.

Los diferentes motores de búsqueda al aplicar algoritmos diferentes, obtienen también resultados diferentes, coincidiendo en la mayoría de asignaciones, pero no en todas. Así pues,

los resultados de una búsqueda sobre un determinado conjunto de espectros con un buscador son complementarios a los obtenidos con ese mismo conjunto de espectros utilizando otro buscador (Figura 6).

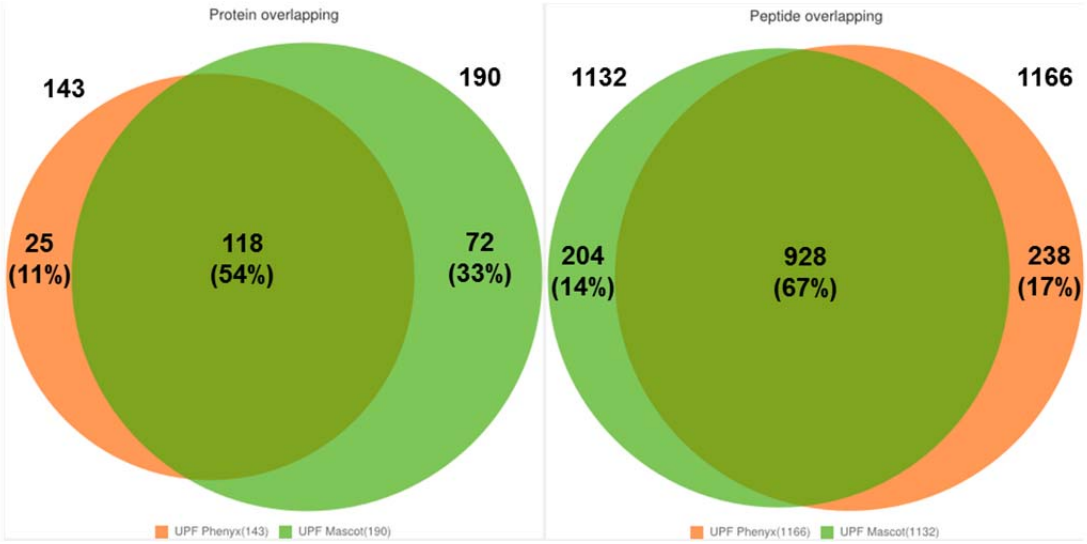


Figura 6. Solapamientos de péptidos y proteínas: Diagramas de Venn mostrando el solapamiento entre los resultados de la búsqueda de un mismo set de espectros utilizando motores de búsqueda distintos, en este caso, Phenyx (naranja) y Mascot (verde). Los datos provienen de un análisis del laboratorio de proteómica de la Universidad Pompeu Fabra (UPF) y los diagramas de Venn se han generado con la herramienta MIAPE Extractor (sección 4.2.5). En este caso existe un 67% de solapamiento a nivel de péptido y un 54% a nivel de proteína.

Los motores de búsqueda tienen en cuenta únicamente unos tipos de iones concretos de todos los posibles. Lo más común es que se consideren los iones de tipo *b* y tipo *y* derivados de los experimentos con fragmentación/disociación inducida por colisión CID. Sin embargo, casi todas las herramientas permiten al usuario elegir otros tipos de iones a considerar (como los iones *a* y los iones *imonio*) (Tabla 1).

Motor de búsqueda	Comparación de espectros	Normalización de espectros	Series de iones considerados	Intensidades consideradas	Tiene en cuenta el ruido	Semi-tríptico	Puntuaciones
SEQUEST	Inter-correlación	Si	*b, y, a* (*neutral losses* en versión comercial)	Implícitamente	Implícitamente	Si	Xcorr y ΔCn
MASCOT	Probabilística	?	Todas	No	Implícitamente	Si	Ion score y e-value
X!Tandem	Producto escalar	Si	*sólo b, y*	No	No	Si	Hyperscore y e-value
OMSSA	Número de picos de iones asignados	No	*b, y* (ECD/ETD también disponible)	No	Explícitamente (quitando los picos menos intensos, y aquellos alrededor de los principales)	Si	e-value
Phenyx	Logaritmo de la probabilidad de la puntuación	No	*Todas*	Explícitamente	No	Si	Logs score, *p* -value y Z-score
MyriMatch	Probabilística	Si	*b, y*	Implícitamente	Explícitamente (usando un % del *total ion current*)	Si	p-value de una asignación aleatoria

Tabla 1. Propiedades de los motores de búsqueda más populares

SEQUEST (Eng, McCormack et al. 1994, Yates, Eng et al. 1995, Yates, Eng et al. 1995) normaliza los espectros teóricos como hace con los experimentales, forzando a que los iones tipo *b* y tipo *y* sean los más intensos. Éste calcula la puntuación de correlación, llamada XCorr, que se acompaña con la puntuación ΔC_n, que mide la diferencia entre la puntuación de la mejor asignación y la segunda mejor (MacCoss, Wu et al. 2002, Eng, Fischer et al. 2008). Aunque habitualmente usa los iones de tipo b y tipo y, SEQUEST es capaz de tener en cuenta cualquiera de las series de fragmentación conocidas.

OMSSA (Geer, Markey et al. 2004) no considera de ninguna manera las intensidades de los picos una vez que los espectros son procesados para quitarles el ruido basal. Basa su puntuación en el número de picos asignados y en una distribución de puntuaciones de Poisson (Fenyo y Beavis 2003). Además de las series b e *y*, es también capaz de procesar las series *c* y *z* de los espectros generados con ETD/ECD.

X!Tandem (Craig y Beavis 2004) define una puntuación *hyperscore* basada en el producto escalar de las intensidades asignadas de las series *b* e *y*. La puntuación es en realidad proporcional a la suma de las intensidades de los iones asignados de las series *b* e *y*.

MyriMatch (Tabb, Fernando et al. 2007) se basa en un principio similar a SEQUEST u OMSSA ya que sus puntuaciones se basan también en la correlación de las masa teóricas y las detectadas experimentalmente, sin considerar explícitamente la intensidad de los iones detectados. Para ello utiliza un modelo multi-modal de intensidades de picos de espectros, el cual estima la probabilidad *p* de una asignación aleatoria. Este *p*-value (ver sección 1.4.3) se transforma posteriormente en una puntuación por medio de la transformación $-ln(p)$.

Por su parte **Phenyx** (Colinge, Masselot et al. 2003, Colinge, Masselot et al. 2004, Magnin, Masselot et al. 2004) usa un modelo probabilístico bastante complejo que tiene en consideración las intensidades de los picos así como todos los tipos de iones. Dicho algoritmo se centra en la probabilidad de observar un PSM por azar basado en varios parámetros: los picos observados que han sido asignados, las series de iones, las intensidades relativas de los picos, las modificaciones post-traduccionales, los errores de las masas de los péptidos y de los productos, los números de corte enzimáticos omitidos y la composición de aminoácidos. A continuación esta probabilidad se transforma en un valor *Z*-value. Este es el número de desviaciones estándares sobre la puntuación media de los péptidos asignados por azar y asume que el logaritmo de las probabilidades de las puntuaciones tiene una distribución normal.

El algoritmo de puntuación de **Mascot** no ha sido publicado, aunque se sabe que es un algoritmo probabilístico que puede tener en cuenta más series que sólo las *b* e *y*, siendo capaz de analizar espectros ETD/ECD. Para cada PSM candidato, se proporciona una puntuación de ion (*ion score*) indicando la probabilidad de que la asignación sea correcta. De hecho, la puntuación

de ion de Mascot es una transformación de esa probabilidad: $-10 \times log_{10}(p)$, de tal manera que una probabilidad de 10^{-20} corresponderá con una puntuación de 200. Mascot también calcula dos puntuaciones de corte: el corte por identidad (*identity)* o por homología *(homology)*. Se estima que el corte por identidad equivale a un 5% de PSMs falsos positivos, siendo levemente superior en el caso del corte por homología.

1.4.3. Validación de resultados de identificación de péptidos: estimación estadística de la confianza

Tras la asignación de cada espectro a una o más secuencias peptídicas que corresponden a espectros teóricos, y el cálculo de la o las puntuaciones asociadas por parte del buscador, obtenemos una lista de péptidos candidatos normalmente ordenados por una de las puntuaciones que se les asigna. Dependiendo del motor de búsqueda, obtendremos unas puntuaciones diferentes, y en la mayoría de los casos, no comparables entre sí. En muchos casos, estas puntaciones asignadas a los PSMs no suelen tener información estadística de la fiabilidad de la asignación y por tanto no son fáciles de interpretar. Además, en la mayoría de buscadores la tasa de error no es conocida.

Nos referiremos a una "identificación" cuando hablamos de una de las asignaciones de la lista ordenada, filtrada por un criterio determinado basado en la puntuación. Sin embargo, los buscadores actuales son incapaces de discernir completamente entre las identificaciones correctas y las incorrectas, de manera que parte de las identificaciones pueden corresponder a falsos positivos y por tanto, siendo estrictos, nos debemos referir a ellas como "identificaciones potenciales". Este sistema de asignaciones debe considerarse imperfecto por varias razones: (1) las proteínas presentes en la muestra no siempre están presentes en el espacio de búsqueda de la base de datos; (2) los espectros derivados de especies no peptídicas presentes en la muestra derivan frecuentemente en asignaciones a secuencias peptídicas; (3) secuencias peptídicas asignadas incorrectamente a veces obtienen puntuaciones mayores que las correctas; (4) tener en cuenta todas las posibles modificaciones post-traduccionales (PTMs) que pueden presentar los péptidos es muy costoso computacionalmente, por lo que sólo un número pequeño de ellas se tiene en cuenta en la búsqueda, y por tanto es probable que otras modificaciones presentes en la muestra dificulten o directamente impidan su identificación. Por tanto, la primera tarea para los investigadores en Proteómica es determinar cuál es la tasa de error que se comete en la lista de asignaciones identificadas.

Distribuciones de puntuaciones

En general, las puntaciones asignadas a un PSM se pueden dividir en dos tipos. Por un lado, las puntuaciones basadas en **factores paramétricos**, que tienen en cuenta exclusivamente las propiedades del espectro real y el teórico para puntuar la asignación. Este es el caso de la puntuación XCorr del buscador SEQUEST. Por otro lado están las puntaciones basadas en **distribuciones**, que estiman la probabilidad de que ese PSM sea correcto o no, para lo cual tienen en cuenta otros PSMs candidatos.

Hay dos tipos de distribuciones; las que describen las asignaciones de cada uno de los espectros experimentales con todos las secuencias peptídicas candidatas a las que se enfrenta, denominadas **distribuciones individuales** de puntuaciones de espectro (*o single-spectrum*), existiendo una para cada espectro experimental, y las que describen las mejores asignaciones de cada espectro dentro del conjunto total de espectros de un experimento dado, denominadas **distribuciones promedio** de puntuaciones (*average-score distribution*).

Las distribuciones individuales permiten evaluar la probabilidad de que un espectro experimental se asigne a una secuencia peptídica de la base de datos, siendo la asignación final estadísticamente significativa con respecto al resto de posibles asignaciones de ese espectro a otras posibles secuencias en la base de datos. Este concepto es semejante al del valor de expectación (*E*-value) (ver apartado de los estadísticos *p*-value y *e*-value), usado con frecuencia por los algoritmos de búsqueda de información en bases de datos, como BLAST o FASTA (Altschul, Madden et al. 1997).

En el caso de las distribuciones promedio de puntuaciones, las mejores puntuaciones, independientemente de cómo sean, son ordenadas de mejor a peor formando un ranking. Utilizadas de esta forma se construye una distribución acumulativa de puntuaciones, que es utilizada para determinar la significatividad estadística de la identificación peptídica. Repitiendo el proceso con una base de datos señuelo, estas distribuciones permiten calcular a su vez el error cometido en el grupo de péptidos identificados. Estas distribuciones acumulativas son estimaciones estadísticas de la distribución promedio de probabilidad de la mejor puntuación, una función que nos indica la probabilidad de encontrar un espectro con puntuación igual o mejor que la puntuación observada, dentro de un experimento dado.

Algunos motores de búsqueda como X!Tandem, Phenyx, o Mascot utilizan una distribución individual formada por la distribución de las puntuaciones de un espectro experimental enfrentado con todos los posibles candidatos en la base de datos. Sin embargo, en la Proteómica de segunda generación, en la que se generan cientos de miles de espectros en un mismo experimento, es esencial conocer también la distribución global de las asignaciones de

todos los espectros en la base de datos para así poder estimar una tasa de error global sobre el número de identificaciones obtenido.

Los estadísticos *p*-value y *E*-value

El valor "***p*-value**" es comúnmente utilizado en bioestadística como medida de la significatividad de los resultados. Definir qué es el "*p*-value" requiere que se defina primero la noción de hipótesis nula. En el caso de la asignación de un péptido a un espectro, la hipótesis nula es, simplificando, que ese espectro no corresponda a ningún péptido de la base de datos. En consecuencia, el *p*-value indica la probabilidad de observar por azar un PSM con una cierta puntación o mayor, asumiendo que se cumple la hipótesis nula. Un valor pequeño de *p*-value indica que la probabilidad de observar un PSM incorrecto es pequeña, y por tanto que el PSM es probablemente correcto.

Cuando el *p*-value se aplica al análisis de la significatividad de distribuciones individuales de espectro, puede inducir a error cuando se aplica al análisis de experimentos de identificación a gran escala. Por ejemplo, dadas 10.000 PSMs con un umbral de puntación asociado a un *p*-value de 0,05, esperamos tener 0,05x10.000=500 PSMs incorrectos sólo por azar. En otras palabras, tendríamos 500 identificaciones incorrectas. Para ello se introduce la corrección de hipótesis múltiple de Bonferroni (Abdi 2007), que sugiere corregir el valor de *p*-value teniendo en cuenta el número de test realizados. Si en un experimento se quiere obtener un valor de significatividad de 0,05 y se repite el test 1.000 veces, entonces se deberá ajustar la puntuación de corte de 0,05 a 0,05/1.000=0,00005. O, dicho de otra manera, la significatividad real del análisis es 1.000 veces menor al tener en cuenta que estamos realizando una hipótesis múltiple que incluye 1.000 hipótesis individuales.

Cuando se analizan distribuciones individuales de espectro para determinar la significatividad estadística de cada espectro de forma independiente, también hay que considerar el efecto de la hipótesis múltiple, porque el PSM corresponde a la mejor puntuación obtenida después de comparar el espectro con un número muy alto de candidatos, lo que equivale a repetir el test estadístico tantas veces como candidatos hemos analizado. Esta consideración nos lleva al concepto de "***E*-value**" o valor de expectación (*Expectation value*), que es un método para la corrección de la hipótesis múltiple en este tipo de análisis. El *E*-value se define como el *p*-value del PSM con mejor puntuación multiplicado por el número de candidatos que se han comparado con el espectro. Los *E*-value se incluyen en los resultados de motores de búsqueda como Mascot, OMSSA o X!Tandem e indican el número de veces que se espera observar el PSM con una cierta puntuación x por azar.

Corrección de la hipótesis múltiple mediante la FDR y las bases de datos señuelo

Una alternativa ampliamente usada para modelar la hipótesis nula es buscar el conjunto de espectros contra una **base de datos señuelo** (*decoy*). Una base de datos señuelo está compuesta por secuencias de aminoácidos derivados o no de la base de datos original, que llamaremos "objetivo" (*target*), pero que en todo caso forman un conjunto de secuencias que interpretamos como falsas y por tanto cualquier asignación de un espectro a cualquiera de estas secuencias asumimos que es una asignación al azar y cumplirá la hipótesis nula.

Por medio de esta aproximación podremos calcular la **tasa de falsos positivos**, (**FDR**, *False Discovery Rate*) (Soric 1989, Benjamini y Hochberg 1995), que se define como la proporción esperada de predicciones incorrectas sobre un conjunto de predicciones. Aplicado a la espectrometría de masas, corresponde con la fracción de PSMs incorrectos existente en un conjunto de PSMs identificados usando un umbral determinado de puntación. Si por ejemplo tenemos 1.000 PSMs con una puntación mayor que una puntación de corte determinada y sabemos que 10 de ellos son PSMs incorrectos, tendremos por tanto una FDR de 1%. Si la búsqueda ha sido realizada con una base de datos señuelo concatenada a una base de datos objetivo (es decir, no han sido dos búsquedas independientes), debemos multiplicar por 2 el valor de la tasa de error, ya que la distribución de falsas asignaciones en la base de datos señuelo se espera que sea igual en la base de datos objetivo (es decir, una distribución aleatoria), y por tanto el número de falsas asignaciones será el doble. La FDR puede ser utilizada para conseguir un compromiso entre el valor de la tasa de error y la sensibilidad, dependiendo de las necesidades y objetivos de cada experimento. Por ejemplo, si queremos obtener una tasa de error de 1%, podremos elegir una puntación de corte determinada para obtenerla y que será diferente de si queremos una tasa de error de 5%.

Asimismo, el uso de una base de datos señuelo nos permitirá también calcular los valores *p*-value asociados a cada PSM de manera directa: para un PSM dado con una puntuación *s*, el valor de p-value será el porcentaje de PSMs señuelo con una puntuación *s* o mejor. Habitualmente se utilizan bases de datos señuelo con el mismo tamaño que las bases de datos objetivo, aunque no es imprescindible. De hecho, cuanto mayor sea la base de datos señuelo, más preciso será el cálculo de la *p*-value, aunque el coste computacional será más elevado (Käll, Storey et al. 2008).

La estrategia de búsqueda en bases de datos señuelo (*decoy*) para calcular la FDR es la más usada en la actualidad por la comunidad proteómica. Por ello ha dado lugar a muchos estudios orientados a analizar y mejorar su fiabilidad (Peng, Elias et al. 2003, Nesvizhskii, Vitek et al. 2007, Choi y Nesvizhskii 2008, Käll, Storey et al. 2008, Käll, Storey et al. 2008, Kim, Gupta et al. 2008, Tabb 2008, Wang, Wu et al. 2009). En el trabajo realizado por P. Navarro y J.

Vázquez (Navarro y Vazquez 2009) se describió un método refinado para el cálculo de la tasa de error basado en la búsqueda separada en las bases de datos señuelo y normal, teniendo en cuenta la equiprobabilidad de resultados de puntuación entre ambas bases de datos para las asignaciones aleatorias.

El estadístico q-value

Para asociar de manera unívoca cada puntación y PSM con su valor de FDR asociado, Storey y Tibshirani propusieron una nueva métrica, el "***q*-value**" (Storey y Tibshirani 2003), que posteriormente introduciría Käll en la espectrometría de masas (Käll, Storey et al. 2008, Käll, Storey et al. 2008). Un valor de *q*-value se entiende como la mínima tasa de error con la que un PSM es aceptado y pasa el corte, o lo que es lo mismo, si tenemos de 0,01 de *q*-value sobre un PSM concreto significa que si aplicamos todas las posibles puntuaciones de corte por FDR, el 1% será la puntuación de corte mínima con la que dicho PSM aparecerá en la lista de identificaciones final. En la práctica, el *q*-value corresponde a la distribución estadística de la FDR.

Aunque los *q*-value se asocian a PSMs concretos, es importante remarcar que los *q*-value dependen del conjunto global de datos en el cual se ha obtenido dicha asignación, es decir, dependen de cada búsqueda individual, ya que para los mismos valores de puntación en dos búsquedas diferentes (con diferentes parámetros o usando una base de datos diferente), obtendremos diferentes valores de *q*-value.

El estadístico PEP

En el caso típico en el que un investigador esté interesado en identificar el conjunto de proteínas que se encuentran presentes en una muestra bajo ciertas condiciones experimentales se utilizará la anteriormente descrita FDR o tasa de error, o los *p*-value para saber qué porcentaje de los PSMs obtenidos son correctos o incorrectos. Sin embargo, es posible otro tipo de escenario en el que lo que se quiere determinar es la presencia o ausencia de un determinado péptido o proteína, por ejemplo, para saber si una proteína se expresa o no en un determinado tipo celular bajo unas determinadas condiciones. Por ejemplo, una vez obtenidos un conjunto de PSMs con una tasa de error baja, nos queremos centrar sólo en uno de ellos y saber su **probabilidad de error a posteriori** (***PEP,*** *Posterior Error Probability*) que se define como la probabilidad de que un PSM sea un falso positivo. Este valor cobra una gran importancia ya que pese a que la tasa de error del conjunto de PSMs obtenidos sea baja, el valor de PEP de un PSM en concreto puede ser bastante alto. Si tenemos ya los valores de PEP de los PSMs, el cálculo de la FDR es trivial, ya que corresponderá a la suma de los valores PEP de los PSMs considerados

significativos, dividido por el número de PSMs significativos. Sin embargo, no es posible el cálculo contrario. La dificultad estriba en que los valores PEP sólo pueden ser calculados conociendo a priori las distribuciones de puntaciones de las asignaciones correctas e incorrectas.

El primer y probablemente más utilizado algoritmo para asignar a los PSMs un valor PEP es el *Peptide-Prophet* (Keller, Nesvizhskii et al. 2002). Este método ajusta un modelo paramétrico a la distribución de puntuaciones de los PSMs de manera no supervisada, lo cual permite luego calcular los valores de PEP para cada uno de los PSMs. Posteriormente, ciertas mejoras en el algoritmo de *Peptide-Prophet* permitieron la inclusión de PSMs señuelo (Choi y Nesvizhskii 2008) y pasar de un modelo paramétrico a uno semi-paramétrico (Choi, Ghosh et al. 2008). Luego, en un trabajo de Käll, Storey et al., (Käll, Storey et al. 2008) se describió un método no paramétrico para obtener los valores PEP, es decir, en el que a priori no se presupone ni el número ni el tipo de parámetros que van a ser necesarios para modelar la distribución de puntuaciones de PSMs, lo cual permite que el método sea más flexible y robusto, pudiéndose aplicar a diferentes tipos de puntuaciones: SEQUEST, Mascot, X!Tandem, InsPect (Tanner, Shu et al. 2005), entre otras.

Validación de resultados de identificaciones para el motor de búsqueda SEQUEST

Como hemos comentado anteriormente, SEQUEST, uno de los primeros y más utilizados buscadores, calcula la puntuación de correlación XCorr y la puntuación ΔC_n, que mide la diferencia entre la mejor y la segunda mejor puntuación XCorr asignada a cada espectro. Estas puntuaciones de los resultados de SEQUEST deben ser posteriormente interpretadas y procesadas para discernir las identificaciones verdaderas de las falsas, evitando así en lo posible los falsos positivos.

Tradicionalmente, los resultados de SEQUEST se empezaron a filtrar de manera empírica, estableciendo un criterio de corte sobre los valores de XCorr y ΔC_n (Link, Eng et al. 1999, Washburn, Wolters et al. 2001, Florens, Washburn et al. 2002, Peng, Elias et al. 2003, Qian, Liu et al. 2005). Sin embargo, estos criterios empíricos no proporcionan la suficiente potencia discriminatoria entre verdaderos y falsos positivos (Keller, Nesvizhskii et al. 2002, Sadygov y Yates 2003, Tabb 2008). Debido a ello, se desarrollaron criterios de filtro alternativos basados en la distribución de las puntuaciones de SEQUEST y/o basados en algoritmos de aprendizaje automático, como el *Peptide-Prophet* para conseguir así una mejor separación entre las asignaciones correctas y las incorrectas (Keller, Nesvizhskii et al. 2002, MacCoss, Wu et al. 2002, Moore, Young et al. 2002, Anderson, Li et al. 2003, Kislinger, Rahman et al. 2003, Lopez-Ferrer, Martinez-Bartolome et al. 2004, Razumovskaya, Olman et al. 2004).

Otros esquemas de puntuación consideran cada uno de los espectros individualmente, en vez de considerar una colección de entera, y tratan de determinar la probabilidad de que la mejor secuencia asignada al espectro se produzca por azar (hipótesis nula). Algunos algoritmos basados en este principio, están construidos sobre un modelo matemático teórico (Bafna y Edwards 2001, Colinge, Masselot et al. 2003, Colinge, Masselot et al. 2004, Geer, Markey et al. 2004), mientras que otros consideran las distribuciones de frecuencias de las puntuaciones obtenidas cuando se busca un espectro contra un conjunto de secuencias candidatas (Fenyo y Beavis 2003).

A pesar de todos estos esfuerzos, el cálculo exacto de la fiabilidad de un péptido identificado se sigue considerando un problema abierto. La falta de indicadores de fiabilidad adecuados es además particularmente problemático cuando se analizan grandes cantidades de datos (Carr, Aebersold et al. 2004) ya que resulta imposible validar manualmente todos los resultados de identificación y existe por tanto una falta de herramientas universalmente aceptadas por la comunidad científica para validar los resultados publicados (Baldwin 2004). De hecho, en el tiempo en el que se hizo este trabajo se consideraba que un gran número (a la par que desconocido) de identificaciones por espectrometría de masas que habían sido ya publicadas seguramente fueran falsos positivos (Keller, Nesvizhskii et al. 2002, MacCoss, Wu et al. 2002), lo que creaba una gran desconfianza en la comunidad científica hacia los resultados encontrados en Proteómica.

Parte del trabajo descrito en esta tesis consiste en un nuevo análisis matemático de las distribuciones promedio de las puntuaciones de SEQUEST, analizando su comportamiento y sus propiedades. Para ello, introducimos el concepto de la calidad del espectro, y relacionamos las distribuciones promedio con las distribuciones individuales de espectro. Usando todos estos conceptos hacemos inferencias sobre la dependencia de las distribuciones promedio del tamaño de la base de datos y finalmente usamos el concepto de la razón de probabilidades como un nuevo indicador estadístico que tiene en cuenta la mejor y la segunda mejor puntuación asignadas a los espectros y que trata de mejoras los problemas anteriormente descritos sobre la discriminación entre verdaderos y falsos positivos.

1.4.4. El problema de la inferencia de las proteínas

En las aproximaciones llamadas "*de abajo arriba*" (*bottom-up)*, se obtienen un conjunto de péptidos de los que posteriormente se infiere qué proteínas pueden estar o no en la muestra. Esta inferencia no es siempre una tarea trivial, incluso partiendo de péptidos identificados con una alta significatividad, y en numerosas ocasiones contienen ciertas ambigüedades con las que hay que saber trabajar. El problema reside principalmente en la redundancia existente en las

secuencias de las proteínas, ya que múltiples isoformas de una misma proteína comparten la mayoría de los péptidos, o bien, secuencias peptídicas con pocos aminoácidos se encuentran frecuentemente en diferentes proteínas, sin necesidad de tener ninguna relación evolutiva. Esta problemática hace que en muchos estudios se identifiquen proteínas que puede que no estuviesen presentes en la muestra y a su vez, puede que se omitan proteínas que realmente sí que estaban presentes en ella. Además, los espectros no asignados a secuencias peptídicas, que suelen ser la mayoría, pueden provenir de secuencias que porten modificaciones post-traduccionales (PTMs), por lo que la presencia de éstas modificaciones específicas de la muestra analizada también dificulta la inferencia de proteínas a partir de los datos experimentales.

En numerosas ocasiones se opta por la inferencia de proteínas basada en el principio de "*La navaja de Occam" ("Occam's razor")* que consiste en buscar la combinación de proteínas más sencilla que explique la presencia de todos los péptidos identificados. Podemos encontrar una detallada revisión sobre el problema de la inferencia de las proteínas en la publicación de A.I. Nesvizhskii y R. Aebersold (Nesvizhskii y Aebersold 2005), en la que se clasifican los diferentes posibles situaciones en los que diversos péptidos están presentes en varias proteínas, y en el que clasifican las proteínas en diferentes tipos de grupos, dependiendo de qué péptidos se comparten entre ellas y se describen las distintas ambigüedades posibles. Una simplificación de esta clasificación la presenta el algoritmo *PAnalyzer* (Prieto, Aloria et al. 2012) cuya implementación también se detalla en el apartado de materiales y métodos (3.2.3).

Otras aproximaciones existentes son: *Protein-Prophet* (Nesvizhskii, Keller et al. 2003), que asigna los péptidos a las proteínas en función de la probabilidad estimada de cada proteína. *IsoformResolver* (Meyer-Arendt, Old et al. 2011) agrupa las proteínas basándose en la información experimental derivada del análisis MS/MS y por otro lado con la información derivada de la digestión teórica de las proteínas de una base de datos, que permite realizar una agrupación de proteínas basada en la información funcional de las proteínas. *PeptideClassifier* (Qeli y Ahrens 2010), además de utilizar información de espectrometría de masas, considera información proveniente de otras fuentes para facilitar la asignación entre péptidos y proteínas, como la información de las bases de datos genómicas.

1.5. Estandarización de datos en Proteómica

En los últimos años, el número y tamaño de los datos publicados que derivan de experimentos proteómicos ha crecido espectacularmente. Este crecimiento es resultado directo del desarrollo en instrumentos, métodos y herramientas bioinformáticas capaces de generar, recoger y analizar grandes cantidades de datos. La comunidad científica tiene por tanto la necesidad de tener todos esos datos a su alcance, para extraer conocimiento de ellos, o para el

desarrollo de nuevas herramientas bioinformáticas de análisis. Sin embargo, cada vez más, la publicación de datos conlleva un problema tanto para los autores, como para los revisores y los lectores, debido a la falta de herramientas dirigidas a recopilar, validar o presentar convenientemente los datos de identificaciones a gran escala (Baldwin 2004).

La mayoría de las revistas especializadas en Proteómica, comenzaron a proponer ciertas directrices (Carr, Aebersold et al. 2004) en las que se detalla la información que debían contener los datos de, por ejemplo, identificaciones a gran escala, así como la existencia de ciertos criterios de control de calidad mínimos que deben cumplir (análisis estadísticos de validación). Estas iniciativas surgieron en parte por el hecho de que un número más o menos grande de las proteínas que estaban siendo publicadas como "identificadas" eran en realidad falsos positivos (Keller, Nesvizhskii et al. 2002, MacCoss, Wu et al. 2002). Sin embargo, seguía habiendo mucha heterogeneidad entre los criterios que las revistas imponían como mínimo e indispensable para publicar.

Para poder utilizar los datos depositados provenientes de cualquier experimento proteómico de cualquier laboratorio, es necesario que dichos datos puedan ser legibles e interpretables, ya que existen numerosos tipos de espectrómetros fabricados por distintas casas comerciales y cuyos datos generados suelen ser por lo general ficheros binarios en formato propietario, inaccesibles sin las librerías o herramientas del fabricante. Es por tanto fundamental utilizar estándares de representación de datos proteómicos que permitan a terceros poder interpretar los datos disponibles y así reanalizar o incluso comparar datos provenientes de diferentes plataformas Proteómicas.

1.5.1. La iniciativa de estandarización de Proteómica en HUPO

En el año 2002 se creó la iniciativa de estandarización en Proteómica (*PSI: Proteomics Standards Initiative*) por parte de la Organización del Proteoma Humano (*HUPO: Human Proteome Project*), esto es, el ***HUPO-PSI***. Durante estos más de diez años ha trabajado en la estandarización de distintos aspectos de los flujos de trabajo en Proteómica.

El HUPO-PSI es una iniciativa en la que tanto investigadores como desarrolladores de software, representantes de casas comerciales, mantenedores de bases de datos biológicas, o representantes de revistas especializadas en Proteómica han participado activamente en el desarrollo y definición de distintos tipos estándares. Es por tanto una iniciativa abierta con una clara intención de trabajar de forma colaborativa para beneficiar a la comunidad científica Proteómica.

HUPO-PSI se divide en distintos grupos de trabajo, cada uno especializado en un área específica de la Proteómica: Interacciones de proteínas (*PI: Protein Interactions*), Separaciones de proteínas (*PS: Protein separations*), Espectrometría de masas (*MS: Mass Spectrometry*) e Informática de las proteínas (*PI: Protein Informatics*).

Las definiciones de los estándares que se desarrollan en esta iniciativa pasan un estricto proceso de revisión coordinado por el editor del PSI (Vizcaino, Martens et al. 2007). Los estándares que pasan dicha revisión son declarados como *estándares finales*, y normalmente son publicados en una revista científica, pasando a su vez el propio proceso de revisión de la revista, considerándose a partir de entonces como estándares aceptados por la comunidad científica Proteómica.

Nuestro grupo de investigación juega un papel muy activo en la iniciativa del PSI, involucrando además a miembros de otros laboratorios de la plataforma ProteoRed: organizó la reunión anual de la iniciativa en Toledo en 2008, y posee actualmente responsabilidades en la coordinación de la iniciativa, y desarrolla diversos roles dentro del grupo de trabajo "*Protein Separation*", así como la coordinación del desarrollo y mantenimiento de las directrices MIAPE del grupo de trabajo "*Proteomics Informatics*" (Tabla 2).

	Chair	Co-chairs	Editors	Minimum Reporting requirements	Ontology	Secretary	Website content
PSI Steering committee	Eric Deutsch	Henning Hermjakob, Pierre-Alain Binz	Martin Eisenacher, Andy Jones	Pierre-Alain Binz	Gerhard Mayer	Sandra Orchard	Sandra Orchard, Martin Eisenacher
Mass Spectrometry Data Interchange (MS)	Eric Deutsch	Pierre-Alain Binz		Pierre-Alain Binz	Matt Chambers		
Molecular Interactions (MI)	Henning Hermjakob	Sylvie Ricard-Blum	Sandra Orchard	Sandra Orchard	Sandra Orchard	Lukasz Salwinski	Rafael C Jimenez
Proteomics Informatics (PI)	Andy Jones	Martin Eisenacher, Juan Antonio Vizcaino	Gerhard Mayer	Salvador Martinez de Bartolomé	Gerhard Mayer	Juan Antonio Vizcaino	Da Qi
Protein Modifications (MOD)	John Garavelli	Sean Seymour		N/A			
Protein Separation (PS)	Juan Pablo Albar	Andy Jones	Salvador Martinez de Bartolomé	Montserrat Carrascal	J. Alberto Medina		Salvador Martinez de Bartolomé

Tabla 2. Asignación de responsabilidades y roles en los distintos grupos de trabajo del HUPO-PSI: Marcados en colores naranjas se destacan los nombres de personas de nuestro grupo o de ProteoRed (en el caso de Montserrat Carrascal) que llevan a cabo una de las responsabilidades en el HUPO-PSI. En naranja más oscuro las responsabilidades de Salvador Martínez de Bartolomé.

1.5.2. Estándares de representación de datos proteómicos

Los estándares para la representación, almacenamiento e intercambio de datos proteómicos pretenden unificar el formato y la información que deben contener los diferentes tipos de datos proteómicos. Estos estándares están mayoritariamente definidos con una estructura en XML (*eXtensible Markup Lenguaje:* lenguaje de marcas extensible) para hacer más fácil su lectura por parte de herramientas bioinformáticas externas, debido al gran número de tecnologías y librerías capaces de trabajar con ellos. Por otra parte, para formalizar la estructura de los ficheros XML existen los llamados ficheros de definición de esquemas XML, esto es, los XSD (*Xml Schema Definition*), que definen qué elementos forman los ficheros XML, y qué referencias existen entre ellos.

Sin embargo, cuando tenemos que representar datos tan heterogéneos como los provenientes de un experimento proteómico, la definición del esquema no es suficiente y se hacen necesarias estructuras adicionales que proporcionen una flexibilidad mínima como para que se puedan albergar dichos datos de manera controlada y entendible. Para ello, los esquemas XML en Proteómica contienen unas estructuras XML llamadas "***cvParam***" y "***userParam***" que permiten definir pares nombre/valor, en el primer caso, utilizando términos de ontologías, y en el segundo, abiertas a cualquier valor:

<cvParam name="*fragment neutral loss*" value="*0*" accession="*MS:1001524*" cvRef="*PSI-MS*" unitName="*dalton*" unitAccession="*UO:0000221*" unitCvRef="*UO*"/>

<userParam name="*Mascot User Comment*" value="*papilomavirus*"/>

Existen diferentes estándares de representación de datos proteómicos dependiendo de la naturaleza de la información que representan, la cual depende de la técnica Proteómica utilizada o de la etapa en el flujo de trabajo de un análisis proteómico. En las siguientes líneas se describirán únicamente los estándares relacionados con el trabajo expuesto en la tesis y el resto únicamente se nombrarán.

El estándar **GelML** (*Gel Electrophoresis Markup Languaje*) (Gibson, Hoogland et al. 2010) representa información sobre experimentos de separación de péptidos y/o proteínas por medio de electroforesis en gel, y puede contener diferentes metadatos (información sobre los protocolos experimentales) y datos (las propias imágenes de los geles) provenientes de dicha técnica. Conjuntamente, se ha desarrollado una ontología de vocabularios controlados para anotar convenientemente todos los datos en el fichero (sepCV).

En cuanto a los datos de salida de los espectrómetros de masas, HUPO-PSI quiso proveer un formato para poder representarlos, siendo éste independiente de cualquier fabricante y sus respectivos formatos propietarios. Así pues, se definió el estándar **mzML** (Deutsch 2008,

Deutsch 2010, Martens, Chambers et al. 2011), con la intención de unificar anteriores esfuerzos en este sentido, como el mzXML (Pedrioli, Eng et al. 2004) definido por el ISB (Institute of Systems Biology, US), y el mzData (Orchard, Taylor et al. 2004) definido inicialmente por el HUPO-PSI. En este caso existe también una ontología con los vocabularios controlados necesarios para anotar la información en los ficheros mzML llamada PSI-MS. Para poder interpretar, visualizar y/o procesar la información contenida en estos ficheros, se han desarrollado diferentes librerías tanto en Java (jmzML) (Cote, Reisinger et al. 2010), en Python (pymzML) (Bald, Barth et al. 2012) o en R, como un paquete de Bioconductor (Reimers y Carey 2006) llamado mzR.

El estándar **mzIdentML** (Eisenacher 2011, Jones, Eisenacher et al. 2012) fue definido para representar datos acerca de la identificación de péptidos y proteínas a partir de espectros de masas. mzIdentML contiene las listas de péptidos y proteínas identificadas en un experimento, además de metadatos que describen los métodos, los parámetros utilizados o las métricas de calidad de dichas identificaciones. De la misma manera que el estándar mzML, utiliza la ontología PSI-MS para anotar convenientemente todos los datos dentro de la estructura XML. Existen numerosas implementaciones de herramientas que utilizan el estándar mzIdentML tanto para leerlos como para generarlos. Algunas de ellas son: el motor de búsqueda Mascot (a partir de la versión 2.3), la herramienta "*Proteomics Conversion Tool (ProCon)*" que convierte los resultados del motor de búsqueda SEQUEST en mzIdentML, el conversor "*idConvert*" del conjunto de herramientas *ProteoWizard* (Kessner, Chambers et al. 2008, Chambers, Maclean et al. 2012) que convierte los ficheros pepXML y protXML en mzIdentML o la librería *jmzIdentML* escrita en Java, para la lectura y escritura de ficheros mzIdentML (Reisinger, Krishna et al. 2012). Además está el recientemente publicado "*toolkit*" para mzIdentML en el que se participó con el validador semántico de ficheros mzIdentML (Ghali, Krishna et al. 2013) (ver sección 4.2.2.1).

Otros formatos a destacar, aunque fuera de la temática del trabajo aquí descrito son el **TraML** (Deutsch, Chambers et al. 2012), que permite representar listas de transiciones y sus metadatos asociados con la intención de incrementar la usabilidad y el intercambio de las transiciones optimizadas, y el recientemente publicado **mzQuantML** (Walzer, Qi et al. 2013) estándar para representar datos cuantitativos de péptidos y proteínas. Por su parte, el grupo de trabajo de interacciones moleculares desarrolló otros estándares entre los que están el **PSI-MI XML**.

1.5.3. Directrices para la descripción de experimentos proteómicos: MIAPEs

El crecimiento de las técnicas de alto rendimiento (*high throughput*) en los últimos 20 años, no sólo en Proteómica, ha creado la clara necesidad de crear directrices sobre cómo presentar o publicar este tipo de experimentos (Carr, Aebersold et al. 2004). En el caso del campo de los microarrays, y antes que en Proteómica, la sociedad de datos de genómica funcional (*FGED - Functional GEnomics Data Society*) definió las directrices MIAME (*Minimum Information About a Microarrays Experiment*) (Brazma, Hingamp et al. 2001) que indicaban la información mínima que se debe proporcionar en este tipo de experimentos. En el campo de estudios de diagnósticos médicos, un grupo de científicos y editores definieron análogamente el estándar llamado STARD (*Statement for Reporting Studies of Diagnostic Accuracy*) (Bossuyt, Reitsma et al. 2003).

En el caso de la Proteómica, en 2004 se definieron las llamadas directrices PARIS, posteriormente revisadas en el 2005 y adoptadas por la revista *Molecular & Cellular Proteomics* en sus instrucciones para autores (Carr, Aebersold et al. 2004). Posteriormente, durante los años 2006 y 2007 HUPO-PSI, definió el concepto de **MIAPE** (***Minimum Information About a Proteomics Experiment***) como la información mínima necesaria para interpretar de forma precisa los resultados de un experimento proteómico y permitir la reproducción del mismo (Taylor 2006, Taylor, Paton et al. 2007). Así pues, seguir las directrices MIAPE significa que además de los resultados del experimento se deben incluir una serie de metadatos adicionales que describen de forma suficiente los protocolos, algoritmos y herramientas utilizadas para obtener dichos resultados.

En ese momento, el PSI junto con el GSC (*Genomic Standard Consortium*) y el MGED RSBI Working Groups (*Reporting Structure for Biological Investigations Working Groups*) iniciaron un proyecto de generación y compilación de directrices a nivel global en las distintas áreas de las ciencias biológicas y biomédicas, el llamado proyecto **MIBBI** (*Minimum Information about Biological and Biomedical sciences*) (Taylor, Field et al. 2008, Kettner, Field et al. 2010), que recopila los estándares desarrollados por las distintas iniciativas de estandarización en el campo de la investigación biológica y biomédica.

Gracias a las directrices MIAPE, los informes o publicaciones de experimentos proteómicos que siguen estas directrices no sólo contendrán los resultados en sí, sino que irán acompañados de un conjunto de metadatos con información acerca del diseño y la metodología de los análisis experimentales o computacionales llevados a cabo en cada una de las fases del experimento, lo que aportará sin duda un sello de calidad a la publicación y a la revista científica en la que se publique dicho trabajo. Este documento sentó inicialmente las bases para los subsiguientes

módulos MIAPE que se fueron desarrollando posteriormente, cada uno describiendo la información mínima a publicar acerca de una parte experimental o del análisis en los que un experimento proteómico puede dividirse.

Las directrices MIAPE no tienen por objetivo recopilar absolutamente toda la información relacionada con un experimento. Su definición se basa en dos criterios principales: 1) **suficiencia**: debe recoger la información suficiente acerca de los datos y del contexto experimental que permitan al lector entender y evaluar el experimento críticamente, y 2) **practicidad**: cumplir las directrices MIAPE no debe impedir o dificultar su uso.

Al igual que con el resto de estándares del HUPO-PSI, pero siendo más relevante en este caso, las directrices MIAPE son el resultado de la discusión de diferentes sectores involucrados en el mundo de la Proteómica: editores de revistas especializadas, desarrolladores de software libre, representantes de casas comerciales que desarrollan también su propio software, mantenedores de bases de datos y, por supuesto, experimentalistas. Gracias a todos ellos se ha conseguido un conjunto de directrices que tras pasar a su vez por el estricto proceso de edición de los estándares del PSI son consideradas como estándares en la amplia comunidad Proteómica internacional. Adicionalmente, y para incrementar su visibilidad, cada uno de los módulos MIAPE desarrollados han sido publicados en revistas científicas especializadas en Proteómica como *Nature Biotechnology* o *Journal of Proteomics*, pasando previamente por los respectivos procesos de revisión de dichas revistas.

Sin embargo, en la práctica son las revistas las que definen qué información es necesaria incluir en un manuscrito o en un material suplementario para su publicación. Las diferentes revistas especializadas en Proteómica han definido sus propias directrices como instrucciones para los autores (Celis 2004, Bradshaw, Burlingame et al. 2006, Wilkins, Appel et al. 2006), y esta iniciativa del PSI ha intentado unificar dichas directrices. Representantes o editores de varias de ellas se interesaron por la aceptación de unas directrices comunes, atendieron a varias reuniones conjuntas con el PSI en las que se les consultó su opinión y experiencias acerca de diversos temas como: si debían pedir o no a los autores que depositen los datos crudos de espectrometría de masas en repositorios públicos; cómo llevar a cabo un control de calidad de datos proteómicos; cómo permitir y facilitar el reprocesado de los datos; o cómo afrontar la adopción de las directrices MIAPE en sus instrucciones para autores, editores y revisores (Orchard, Hermjakob et al. 2005, Orchard, Hermjakob et al. 2006, Orchard, Binz et al. 2009, Orchard y Ping 2009).

Los diferentes módulos MIAPE que se han definido en el HUPO-PSI son los siguientes (detallados en la Tabla 3):

- **MIAPE Gel Electrophoresis (MIAPE-GE)** (Gibson, Anderson et al. 2008): describe el protocolo experimental por el cual una muestra se somete a una separación mono o bi-dimensional por electroforesis en una matriz. Esto incluye información acerca de la preparación del gel, las condiciones de la carrera electroforética, técnicas de tinción para la visualización de bandas o manchas en el gel e incluso el método de escaneo llevado a cabo para digitalizar los geles.

- **MIAPE Gel Informatics (MIAPE-GI)** (Hoogland, O'Gorman et al. 2010): describe el proceso por el cual las imágenes digitalizadas de los geles son analizados por un software de análisis de imagen para detectar y cuantificar las manchas o las bandas. Esto incluye la descripción del diseño experimental, es decir, la descripción de los distintos grupos y/o réplicas, además de la descripción del software y los parámetros utilizados en él, así como los métodos de análisis llevados a cabo como alineamiento de imágenes, la detección, la asignación y cuantificación de las manchas, o los análisis estadísticos necesarios para determinar la fiabilidad de los resultados de expresión diferencial obtenidos.

- **MIAPE Mass Spectrometry (MIAPE-MS)** (Taylor, Binz et al. 2008): describe el proceso por el cual una muestra es analizada en un espectrómetro de masas para generar unos ficheros de datos crudos (espectros sin procesar), así como el proceso por el cual dichos datos crudos se procesan para generar unos espectros aptos para utilizar como entrada en los motores de búsqueda. Además describe el equipamiento utilizado, incluyendo su configuración y los parámetros para la adquisición de espectros de masas.

- **MIAPE Mass Spectrometry Informatics (MIAPE-MSI)** (Binz, Barkovich et al. 2008): describe el proceso por el cual los espectros de masas adquiridos en el espectrómetro son analizados para identificar proteínas y péptidos existentes en la muestra. Esto incluye la descripción de todo el proceso del motor de búsqueda, esto es, descripción del software y sus parámetros, la descripción de cualquier análisis de secuenciación *de-novo* realizado o de cualquier análisis estadístico o reprocesamiento sobre los datos de identificación obtenidos por el motor de búsqueda.

- **MIAPE Mass Spectrometry Quantification (MIAPE-Quant)** (Martinez-Bartolome, Deutsch et al. 2013): describe todos los procesos experimentales llevados a cabo por un amplio abanico de técnicas de cuantificación sobre datos de espectrometría de masas, como técnicas basadas en marcaje metabólico (SILAC) o químico (iTRAQ, TMT, ICPL, …), así como técnicas sin marcaje (conteo de espectros, alineamiento de cromatogramas, etc…) o técnicas de cuantificación dirigida (SRM/MRM). Este módulo de directrices MIAPE ha sido el último definido por HUPO-PSI, naciendo de la iniciativa de un grupo de expertos de ProteoRed, e incluyéndose finalmente como parte del trabajo mostrado en esta tesis.

Otros módulos desarrollados son: **MIAPE Capillary Electrophoresis (MIAPE-CE)** (Domann, Akashi et al. 2010), **MIAPE Column Chromatography (MIAPE-CC)** (Jones, Carroll et al. 2010), **Minimum Information about a Molecular Interaction experiment (MIMIx)** (Orchard, Salwinski et al. 2007), **Minimum Information About a Protein Affinity Reagent (MIAPAR)** (Bourbeillon, Orchard et al. 2010) y **Minimum Information About a Bioactive Entity (MIABE)** (Orchard, Al-Lazikani et al. 2011).

MIAPE GE	***MIAPE GI***	***MIAPE MS***	***MIAPE MSI***	***MIAPE Quant***
1. Características generales del experimento (datos de contacto, tipo de electroforesis). 2. Descripción de la muestra. 3. Componentes de la matriz del gel y protocolo de electroforesis. 4. Proceso entre la primera y segunda dimensión (equilibrado, reducción, alquilación,…). 5. Proceso de detección (tinción…). 6. Adquisición de imagen (escáner y software de escaneo). 7. Imágenes (formato y disponibilidad).	1. Características generales del experimento (datos de contacto, tipo de experimento de electroforesis, imágenes de entrada del análisis, software,…). 2. Diseño experimental del análisis de los geles (réplicas, grupos, estándares,…). 3. Preparación de las imágenes (escalado, redimensionado, cortes,…). 4. Procesado de imágenes (alineamiento,…). 5. Extracción de datos (detección de spots, macheo, cuantificación de spots,…). 6. Análisis de los datos (determinación de los nivel de expresión significativos,…). 7. Resultados de los análisis.	1. Características generales del experimento (datos de contacto, espectrómetro de masas). 2. Fuentes de ionización (ESI, MALDI, otros,…). 3. Componentes principales tras la fuente de ionización. 3.1. Analizadores. 3.2. Activación / Disociación. 4. Generación y anotación de los espectros y listas de picos. 4.1. Adquisición de los datos (software y parámetros de adquisición). 4.2. Análisis de los datos (software de generación de listas de picos o de espectros procesados). 4.3. Datos resultantes (cromatogramas, datos crudos y/o listas de picos).	1. Características generales del experimento (datos de contacto, software de análisis). 2. Datos de entrada (tipos y formatos) y parámetros del software (tolerancias, missed-cleavages, PTMs, base de datos,…). 3. Salida del análisis 3.1. Información de proteínas (código de acceso, descripción, puntuaciones, cobertura de secuencia,…). 3.2. Información de péptidos (secuencia, puntuaciones, modificaciones, referencia al espectro,...). 4. Interpretación y validación de los resultados (determinación de la significatividad de las identificaciones, análisis estadísticos,…).	1. Características generales del experimento (datos de contacto, aproximación cuantitativa). 2. Diseño experimental y descripción de la muestra (grupos, réplicas, marcaje,…). 3. Datos de entrada (tipos, formatos, agregación, disponibilidad,…). 4. Protocolo de cuantificación (software y algoritmos). 5. Datos cuantitativos resultantes del análisis (en los distintos niveles de agregación).

Tabla 3. Esquemas de la información requerida por las directrices MIAPE GE, GI, MS, MSI y Quant.

1.5.4. Formatos XML, Vocabularios Controlados y directrices MIAPE

Como hemos comentado, los ficheros estándares utilizan las estructuras XML *userParam* y *cvParam* para poder anotar la mayoría de la información que contienen. En estas estructuras, se utilizan pares de nombre-valor para describir características concretas de los protocolos, herramientas, parámetros, resultados, etc. Sin embargo, hace falta un mecanismo para controlar la semántica de estas estructuras y definir qué vocabularios controlados (estructuras *cvParam*), pueden, deben o tienen que estar presentes en los diferentes elementos de un fichero XML estándar.

El sistema de control semántico definido por el PSI (*PSI validator framework*) (Montecchi-Palazzi, Kerrien et al. 2009) se basa en la definición de un conjunto de reglas de correspondencia entre los diferentes elementos del fichero XML estándar y ciertos conjuntos de vocabularios controlados. Estas relaciones, definidas por otros ficheros XML, llamados **ficheros de mapeo de vocabularios controlados,** permiten definir qué conjunto de términos está permitido en cada uno de los elementos definidos en el esquema y con qué severidad (sugerencia->*MAY*, conveniencia->*SHOULD* u obligatoriedad->*MUST*), para que tenga validez semántica. Así pues, todos los estándares de representación de datos definidos por HUPO-PSI están acompañados por un fichero de mapeo de vocabularios controlados, que junto con la definición del esquema y cierta documentación constituye la especificación de dicho estándar.

Como ha sido comentado anteriormente, existe una estrecha relación entre los diferentes módulos de las directrices MIAPE y los formatos de representación de datos proteómicos (Tabla 4). Idealmente, los ficheros estándares deben contener toda la información definida en su correspondiente directriz MIAPE, sin embargo esto es algo que en la realidad sólo ocurre en el caso de algunos estándares, o más bien, en el caso de algunas herramientas generadoras de ficheros estándares. Para poder asegurar que los ficheros estándares están generados de acuerdo a las directrices MIAPE, se deben definir nuevas reglas de validación más estrictas que las descritas anteriormente para la validación semántica. Parte del trabajo de esta tesis descrito a continuación se centra en la definición de estas reglas y la ampliación de las herramientas de validación semántica existentes para soportar la validación MIAPE (Ghali, Krishna et al. 2013).

Grupo de trabajo PSI	Módulo MIAPE	v.	Formato	v.	Ontología
Molecular Interactions	Molecular Interactions (MIMIx)	1.1.2	PSI-MI XML	2.5.4	PSI-MI CV
	Bioactive Entity (MIABE)	1.0.0			
	Protein Affinity Reagent (MIAPAR)	1.0.0	PSI-PAR	1.0.0	PAR CV
Mass Spectrometry	Mass Spectrometry (MIAPE-MS)	2.98	mzML	1.1.0	PSI-MS CV
			traML	1.1.0	
Proteomics Informatics	Identification (MIAPE-MSI)	1.1	mzIdentML	1.1.0	
	Mass Spectrometry Quantification (MIAPE-Quant)	1.0	mzQuantML	1.0.0	
Protein Separations	Gel Electrophoresis (MIAPE-GE)	1.4	gelML	1.1.0	sepCV
	Gel Informatics (MIAPE-GI)	1.0			
	Column Chromatography (MIAPE-CC)	1.1	spML	1.0.0	
	Capillary Electrophoresis (MIAPE-CE)	0.9.3			
	Phosphoproteomics (MIASSPE)	0.9			

Tabla 4. Relación entre los diferentes módulos MIAPE, los estándares de representación de datos y las ontologías, dentro de cada uno de los grupos de trabajo del HUPO-PSI.

1.6. Repositorios de datos proteómicos

La diseminación de datos científicos es una parte integral de la ciencia. En el campo de la Proteómica han aparecido en los últimos años distintas iniciativas para proveer los medios necesarios para compartir públicamente las cada vez más grandes cantidades de datos que se producen y publican diariamente. Es por ello por lo que diferentes repositorios han aparecido para poder almacenar los diferentes datos producidos por la comunidad Proteómica.

1.6.1. UniProtKB

La base de datos UniProt Knowledgebase (UniProtKB) es el principal recurso donde se recopila la información funcional de proteínas, con ricas anotaciones, precisas y consistentes. Además de capturar la información mínima requerida para una entrada en la base de datos UniprotKB (secuencia de aminoácidos, nombre y descripción de la proteína, datos taxonómicos y citas relacionadas), se añaden tantas anotaciones como estén disponibles para cada entrada, incluyendo términos de ontologías, referencias a bases de datos externas, e indicadores de la calidad de la anotación que indican el tipo de evidencia con la que se ha obtenido dicha información.

UniprotKB está formada por dos partes: una que contiene entradas manualmente anotadas de información extraída de la literatura o de análisis computacionales revisados y evaluados, y otra sección con entradas analizadas computacionalmente a la espera de una revisión manual. Dichas secciones son “UniprotKB/Swiss-Prot” y “UniprotKB-TrEMBL”, respectivamente.

1.6.2. Tranche y NCBI Peptidome

El llamado repositorio Tranche empezó como el proyecto de tesis de Jayson Falkner, de la Universidad de Michigan, en 2005. La primera versión de Tranche fue presentada en el congreso de la Asociación Americana de Espectrometría de Masas (*ASMS*) de 2006, en Seattle, Washington. Fue creado con el objetivo de proveer un software que permitiese de manera sencilla el almacenamiento y la diseminación de ficheros de forma segura, escalable, citable y permanente. Tranche contaba con un sistema distribuido de almacenamiento, con el cual se aseguraba la integridad de todos los ficheros allí almacenados. Además permitía la subida de cualquier tipo de fichero y de cualquier tamaño. Así pues, dada su gran versatilidad y facilidad de uso, pronto se convirtió en el principal recurso existente para almacenar e intercambiar los ficheros de datos crudos que cada vez se generaban con mayor tamaño (Hill, Smith et al. 2010, Smith, Hill et al. 2011).

Por otro lado, el repositorio Peptidome, del Centro Nacional de Información Biotecnológica de Estados Unidos (NCBI), se erigió en 2009 como un repositorio para el almacenamiento de datos de identificación de péptidos y proteínas por espectrometría de masas es tándem (Slotta, Barrett et al. 2009, Ji, Barrett et al. 2010).

Sin embargo, actualmente ambos repositorios ya no se encuentran disponibles (Martens 2013) por falta de financiación lo que ha hecho imposible su soporte. Todos los datos contenidos en Peptidome se encuentran disponibles en un servidor FTP del NCBI, y además fueron migrados por completo al repositorio PRIDE (Csordas, Wang et al. 2013). En el caso de Tranche, la situación ha sido más dramática, ya que parece ser que algunos servidores distribuidos han dejado de dar servicio y ciertos datos han dejado de estar disponibles.

1.6.3. PRIDE

PRIDE (*Proteomics Identification Database*) es sin duda el repositorio más utilizado desde hace años para el almacenamiento de los datos de identificación de péptidos y proteínas por espectrometría de masas. PRIDE nació también como el proyecto de tesis de Lennart Martens, quien diseñó un esquema XML para almacenar datos de identificaciones por espectrometría de masas, el cual sigue en uso tras los años (Martens, Hermjakob et al. 2005). Actualmente el repositorio PRIDE se mantiene en el Instituto Europeo de Bioinformática (EBI), en Inglaterra, perteneciente al laboratorio de biología molecular europeo (EMBL) y contiene más de 28.000 experimentos, conteniendo un total de 12 millones y medio de proteínas, 70 millones de secuencias peptídicas y más de 360 millones de espectros.

El equipo PRIDE del EBI ha desarrollado también las herramientas necesarias para el envío de los datos al repositorio, como los PRIDE Converter 1 (Barsnes, Vizcaino et al. 2009, Jones y Martens 2010, Barsnes, Vizcaino et al. 2011) y 2 (Cote, Griss et al. 2012) que permiten convertir diferentes tipos de ficheros en formato PRIDE XML para luego poder subirlo al repositorio. Adicionalmente también han desarrollado herramientas de consulta de los datos contenidos en el repositorio como la versión BioMart (Smedley, Haider et al. 2009) para PRIDE, un entorno web para realizar consultas por medio de la definición de filtros sobre determinadas características de interés de los datos, o el PRIDE Inspector (Wang, Fabregat et al. 2012), una herramienta de uso local que permite la visualización gráfica de las identificaciones y de los espectros asociados de los experimentos almacenados en el repositorio o de ficheros PRIDE XML locales.

1.6.4. PeptideAtlas

PeptideAtlas (Desiere, Deutsch et al. 2005, Desiere, Deutsch et al. 2006, Deutsch 2010, Farrah, Deutsch et al. 2011) es un repositorio creado alrededor del 2005 por el Seattle Proteome Center (SPC) con datos de identificaciones de péptidos y proteínas por espectrometría de masas. Tiene la particularidad que todos sus datos han sido procesados y validados por el ya mencionado Trans Proteomics Pipeline (Keller y Shteynberg 2011), para así asegurar su fiabilidad y calidad. Contiene datos provenientes de diversos organismos, como humano, ratón o levadura, entre otros. Además, a través del SRM Atlas (Picotti, Lam et al. 2008, Huttenhain, Surinova et al. 2013, Picotti, Clement-Ziza et al. 2013), se almacenan y pueden consultar numerosos experimentos de espectrometría de masas dirigida por SRM (*Selected Reaction Monitoring*) en los que se identifican y cuantifican péptidos y proteínas.

1.6.5. ProteomeXchange

ProteomeXchange empezó en el 2011 con el objetivo de proveer una entrada común para el almacenamiento de datos proteómicos y así promover y facilitar su intercambio y diseminación (Hermjakob y Apweiler 2006). Una vez que los datos se depositan, éstos son compartidos por otros repositorios (Figura 7), de manera análoga a la sincronización de secuencias entre las bases de datos EMBL, NCBI y DDBJ. En este caso, los datos se depositan directamente en el repositorio PRIDE, el cual al recibir los datos envía un mensaje al resto de repositorios indicando que un nuevo conjunto de datos se ha incorporado al sistema. Por su parte, PeptideAtlas al recibir dicho mensaje, reprocesará dichos datos con su flujo de análisis del TPP y volverá a depositar otro conjunto de datos, éste ya etiquetado como reprocesado. Por otra

parte, repositorios externos al consorcio, como GPMDB (*Global Proteome Machine Database*) (Beavis 2006, Fenyo, Eriksson et al. 2010) y otros, podrán suscribirse a este mensaje de aviso y reprocesar también los datos, o incorporarlos en sus propios repositorios. La base de datos UniprotKB, la base de datos de referencia para secuencias proteicas, tiene por intención recoger los datos depositados en ProteomeXchange para así incorporar evidencias a nivel de proteína en sus entradas de SwissProt. Los usuarios tendrán también por tanto un único portal de acceso a todos estos datos.

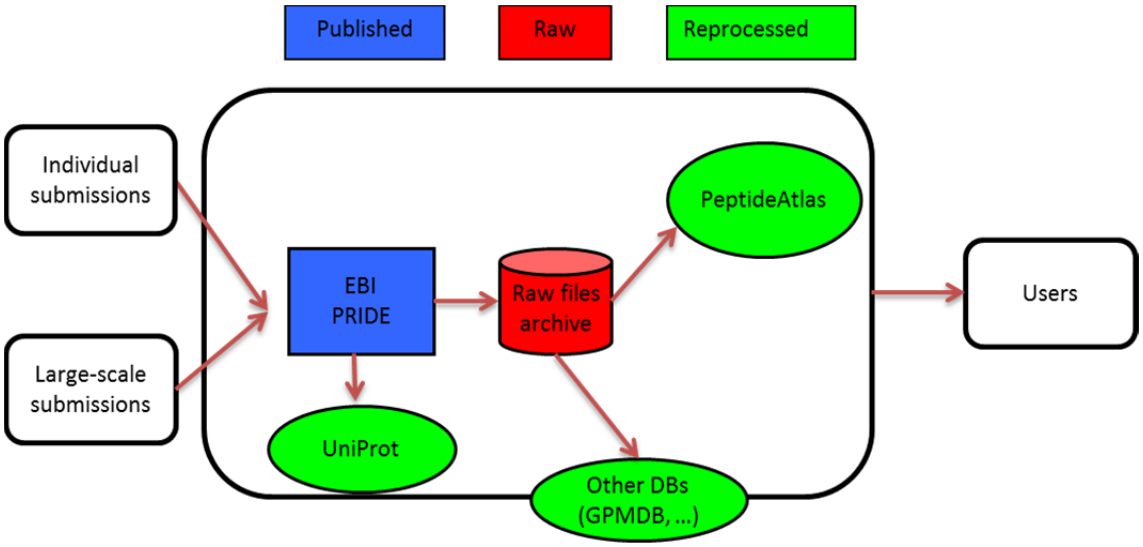


Figura 7. Esquema de flujo de datos en ProteomeXchange. Tanto los envíos de datos individuales como envíos de datos a gran escala entran por una entrada común a ProteomeXchange y los datos se almacenan en el repositorio de EBI PRIDE, incluyendo los archivos de datos crudos (RAW). Estos datos crudos son reprocesados de forma automática por el PeptideAtlas. UniProtKB también recopila la información de espectrometría de masas depositada en el repositorio PRIDE. Otras bases de datos, como GPMDB podrán asimismo reprocesar los datos almacenados en el consorcio.

Cada vez son más comunes los proyectos de investigación colaborativos en los que diferentes laboratorios realizan parte del trabajo, incluso se realizan réplicas técnicas o biológicas de los análisis realizados por el resto de laboratorios con la intención de corroborar los resultados obtenidos. Para la comparación o agregación de los datos proteómicos provenientes de diferentes fuentes es evidente la necesidad del uso de estándares, y en este caso, los estándares desarrollados por la iniciativa de estandarización de HUPO nos proporcionan el "lenguaje común" necesario para la interpretación, el intercambio y la comparación de los datos. La iniciativa HUPO-PSI, así como otros proyectos internacionales como ProDac (*Proteomics Data Collection*, http://www.fp6-prodac.eu) o el propio ProteomeXchange, han realizado un gran esfuerzo en el desarrollo de librerías que permiten el manejo de estos estándares en herramientas bioinformáticas externas. Sin embargo, son realmente pocos los software

desarrollados hasta ahora en proteómica que permiten un flujo completo de análisis basado en dichos estándares. Por otro lado, otro aspecto a considerar con respecto a los estándares en proteómica es su adecuación a las directrices MIAPE, esto es, si tienen la información que se considera mínima para poder interpretar o reanalizar los datos. Actualmente es evidente, sobre todo en el caso de los datos de espectrometría de masas contenidos por el estándar mzML, que dicha adecuación es mínima (ver Tabla 6 más adelante).

Para tratar de resolver o al menos mejorar estos problemas existentes, parte del trabajo de esta tesis se ha enfocado al desarrollo de herramientas bioinformáticas en proteómica basándose en los estándares de HUPO-PSI como: un repositorio de documentos MIAPE (sección 4.2.1), herramientas de validación semántica de la información contenida en los estándares (sección 4.2.2), el desarrollo de una librería para la extracción y manejo de la información MIAPE desde los estándares (sección 4.2.3), así como para el acceso programático al repositorio (sección 4.2.4) y el desarrollo de una herramienta para la integración, la visualización, el análisis y el envío de datos de identificación de péptidos y proteínas a ProteomeXchange, cumpliendo con las directrices MIAPE.

Objetivos

2. Objetivos

Los avances tecnológicos experimentados en los últimos años en la Proteómica permiten obtener cada vez mayores cantidades de datos en un solo experimento. Esta característica común a las tecnologías "ómicas" o de alto rendimiento, ha generado la necesidad de, por un lado, desarrollar algoritmos y herramientas estadísticas para validar el conjunto de identificaciones a gran escala de péptidos y proteínas, minimizando el número de falsos positivos de manera robusta, y por otro lado, desarrollar herramientas que nos permitan integrar, validar, comparar, representar y compartir enormes cantidades de datos provenientes de uno o varios experimentos proteómicos.

Los objetivos principales de esta tesis se enmarcan en dar respuesta a estas necesidades y se concretan en:

1. Desarrollo de algoritmos para la validación estadística de identificaciones por espectrometría de masas en experimentos a gran escala.
   1.1. Definición de un modelo estadístico para el desarrollo de métodos para la validación de identificaciones a gran escala de péptidos y proteínas por espectrometría de masas y su verificación con datos experimentales.
   1.2. Implementación de estos métodos en herramientas informáticas de usuario.
2. Desarrollo de herramientas informáticas basadas en estándares HUPO-PSI:
   2.1. Desarrollo de un repositorio online de experimentos proteómicos basados en las directrices MIAPE.
   2.2. Desarrollo de herramientas de validación semántica y MIAPE de estándares de representación de datos.
   2.3. Desarrollo de una librería para la extracción y el manejo de la información MIAPE.
   2.4. Desarrollo de un acceso programático al repositorio de experimentos proteómicos.
   2.5. Desarrollo de una herramienta informática para proporcionar un flujo completo de integración, análisis, informe y envío automatizado de datos a repositorios internacionales que cumplan las directrices MIAPE en todos sus niveles.
   2.6. Definición de las directrices MIAPE para experimentos cuantitativos en Proteómica, dentro del marco de trabajo del HUPO-PSI.

Materiales y métodos

3. Materiales y métodos

3.1. Validación estadística de resultados de identificación a gran escala

3.1.1. Preparación de muestras y adquisición de datos por espectrometría de masas

Los datos utilizados para la validación estadística de las identificaciones se recogieron de una publicación anterior del grupo (Lopez-Ferrer, Martinez-Bartolome et al. 2004). El primer conjunto de datos consiste en un extracto de proteínas de núcleo de células Jurkat (línea celular derivada de linfocitos T humanos), obtenidos según se describe previamente en (Armesilla, Lorenzo et al. 1999). Brevemente, 100 µg de proteínas provenientes de un extracto celular se precipitó con acetona y se liofilizó a sequedad. Los puentes disulfuro de los residuos de cisteínas fueron reducidos con DTT 10 mM durante una hora en tampón bicarbonato de amonio 25 mM, pH 8 conteniendo urea 8 M, y seguidamente alquilados con iodo acetamida 50 mM durante 45 minutos en condiciones de oscuridad. La mezcla se diluyó 4 veces para reducir la concentración de urea y se sometió a digestión con tripsina (Promega) a 37 C durante toda la noche empleando una relación enzima:sustrato de 1:50. Los péptidos resultantes se separaron por cromatografía de intercambio catiónico seguida por una separación on-line en fase reversa acoplada a una trampa iónica lineal LCQ-DECA XP (Thermo Finnigan). Para el intercambio catiónico se emplearon columnas 0.18x150 mm BioBasic SCX (ThermoHypersil-Keystone) en un sistema microHPLC Smart (Pharmacia) y los péptidos fueron eluídos con un gradiente de KCl de 0 a 122 mM de en tampón fosfato 5 mM, pH 3, en presencia de 25 % acetonitrilo (ACN). Para la fase reversa se emplearon columnas 0.18x150 mm BioBasic C-18 RP (ThermoHypersil-Keystone) conectadas a un sistema HPLC Surveyor (Thermo Finnigan), eluyendo los péptidos con un gradiente de 0 a 48 % de ACN en presencia de ácido acético al 0.5%. El espectrómetro de masas se programó para realizar a lo largo de todo el gradiente cromatográfico la fragmentación MS^2 de los 3 iones más intensos de un barrido desde 400 hasta 1.600 amu (8 µscans, 200 ms IT, 10.000 AGC), empleando exclusión dinámica.

El segundo conjunto de datos utilizado para la validación estadística consiste en un extracto de proteínas de células madre mesenquimales procedentes de médula ósea humana como se describe en una publicación anterior (Ogueta, Munoz et al. 2002).

Un tercer conjunto de datos fue utilizado también, en este caso procedente del análisis de un proteoma de la levadura *Escherichia coli,* proporcionado por la Dra. Michaela Scigelova (Thermo Fisher).

Los tres conjuntos de datos anteriormente descritos fueron adquiridos por medio de diferentes equipos Thermo: 40.000 espectros MS/MS con un LCQ-DECA XP (Thermo Fisher) del análisis de los extractos de células Jurkat, 150.000 espectros MS/MS con un LTQ (Thermo Fisher) del extracto de células madre y 13.000 espectros MS/MS con un LTQ-Orbitrap del proteoma de *Escherichia coli*. El software utilizado para la adquisición de espectros en todos estos casos fue software propietario de Thermo, el Xcalibur versión 2.0.

3.1.2. Motores de búsqueda y bases de datos

Los espectros utilizados en los métodos de validación estadística de las identificaciones fueron analizados utilizando el motor de búsqueda SEQUEST v27, motor de búsqueda incluido en el paquete comercial Bioworks 3.2 (Thermo Electron, CA, USA). Las bases de datos utilizadas para identificar las secuencias peptídicas fueron Uniprot SwissProt (la versión más anotada de Uniprot KB) y NCBI nr (*non-redundant*), con la versión más reciente en cada momento del estudio. Para las muestras humanas (células Jurkat y médula), se aplicó un filtro por taxonomía humana, y para las muestras de levadura, se utilizó la base de datos humana junto con la de levadura para aumentar así la potencia estadística de los resultados.

Las bases de datos aleatorias de secuencias de proteínas, se construyeron usando un programa implementado en Visual Basic 6.0 Edición Empresarial, que invierte la secuencia de cada péptido con la intención de mantener la frecuencia de aminoácidos y la distribución de masas de los péptidos, así como las homologías de secuencia de la base de datos original. La generación de péptidos trípticos por Monte Carlo se realizó también con un programa en Visual Basic que genera aleatoriamente los aminoácidos de acuerdo con su frecuencia natural en las bases de datos.

Los parámetros utilizados en las búsquedas son: 2,0 Da (dalton) de tolerancia en las masas parentales, 1,2 Da de tolerancia para las masas de los fragmentos y modificaciones: Met (+15,9949) variable y Cys (+57,0513) fija.

Los resultados obtenidos por el motor de búsqueda SEQUEST fueron procesados a partir del formato te texto plano .out. Posteriormente a los trabajos aquí descritos, una nueva versión del software que implementaba el método de validación permitió la lectura de ficheros binarios RAW y MSF de Thermo.

3.1.3. Programación y hardware

La implementación del método de la razón de probabilidad se hizo con rutinas de Microsoft Visual Basic 6.0 Edición Empresarial. Todo el trabajo se ha desarrollado bajo la plataforma Windows 2000 Server en un PC Dell Power Edge 1600SC con dos procesadores Intel Xeon 1.8Ghz.

3.1.4. Tasa de error

A lo largo de este trabajo, el término "validación estadística de resultados" de identificación es referido a la determinación de la tasa de error asociada a una lista de péptidos y proteínas identificadas a partir de un conjunto de espectros. Esta tasa de error (*False Discovery Rate*, FDR) indica la proporción en porcentaje de asignaciones de espectros incorrectas dentro de la lista de espectros asignados (PSMs) resultante. Por ejemplo, si en una lista de 10.000 péptidos tenemos asociada una tasa de error del 1%, significa que se estima que 100 de esos 10.000 péptidos son ser falsos positivos. Normalmente, las listas de péptidos identificados se publican ordenados de tal manera que el primero en la lista corresponde con la asignación secuencia-espectro mejor de acuerdo a un parámetro que indica la calidad de cada asignación de la secuencia al espectro. En nuestro trabajo, el parámetro utilizado para ordenar la lista será el llamado **razón de probabilidad** (**pRatio**, *probability Ratio*), cuyo cálculo se describe más adelante (sección 4.1.3). Dentro del orden de la lista de péptidos, podemos asociar a cada uno de los espectros identificados un valor de FDR, que se refiere entonces al FDR máximo que tendría la lista de espectros identificados si el último espectro de la lista fuera aquel al que nos referimos. Este valor de FDR es lo que se describe como *q*-value (Storey y Tibshirani 2003).

3.1.5. El método de la razón de probabilidad

El método de la razón de probabilidad es un método desarrollado sobre un estudio matemático realizado por el Dr. Fernando Martín Maroto. En este trabajo se lleva a cabo la demostración práctica de los conceptos y formulaciones teóricas, así como el diseño, desarrollo e implementación de dos aplicaciones bioinformáticas que permiten su aplicación práctica en estudios de Proteómica realizados en el laboratorio mediante cromatografía bidimensional en tándem con trampa iónica.

Las distribuciones de puntuaciones de probabilidad de espectros individuales se construyeron tras realizar exhaustivas búsquedas sobre secuencias peptídicas aleatorias generadas por Monte Carlo, teniendo en cuenta la frecuencia natural de los aminoácidos en la base de datos NCBI-nr (*non-redundant*), o provenientes de bases de datos inversas, esto es,

bases de datos señuelo. Las distribuciones promedio (*average score distributions*) de los resultados de identificaciones de SEQUEST fueron construidas a partir del parámetro o puntuación XCorr. Las distribuciones promedio de la mejor, de la segunda mejor y de la j-ésima mejor puntuación de SEQUEST fueron construidas a partir de la colección total de espectros, considerando cada una de esas puntuaciones de forma independiente del espectro del que derivan. En la práctica, estas distribuciones de puntuaciones promedio se consiguen tras ordenar todos los datos en función del XCorr, y expresando la probabilidad promedio asociada a cada puntuación como la posición relativa que ocupa en el ranking, esto es (*rank*/E, siendo *rank* la posición que ocupa en el ranking, y *E* el número total de espectros).

Las puntuaciones obtenidas a partir de un mismo espectro pero asumiendo un estado de carga diferente fueron considerados como espectros diferentes.

Únicamente se analizaron espectros con estados de carga 2 o 3, ya que la gran mayoría de los péptidos ionizados mediante ESI tienen uno de estos dos estados de carga.

3.1.6. Aspectos técnicos y metodológicos de las implementaciones de los métodos

Como se mostrará en el apartado de los resultados, tanto el método de la razón de probabilidad como el método de la calidad única, se implementaron en primer lugar en dos herramientas diferentes desarrolladas en Visual Basic 6.0.

La entrada de ambos métodos es el conjunto de ficheros ".out" que proporcionaba SEQUEST.

Como el análisis de un proteoma entero podía suponer ya entonces más de 100.000 asignaciones de espectro, se optó por crear en ambas implementaciones un fichero que unificase la información de los ficheros ".out" de salida de SEQUEST. Una vez creado para cada conjunto de espectros analizados, no era necesario volver a leer todos y cada uno de los ficheros de salida de SEQUEST, resultando bastante más rápido el proceso.

En el caso de la implementación del método de la razón de probabilidad, el fichero con la información conjunta de todos los ".out" se llama "s*core file*". Este fichero almacena la información de la primera y de la segunda mejor puntuación para cada espectro. Una vez leída esta información, el programa forma la distribución de $\mathbf{I_N(x)}$ a partir de las puntuaciones obtenidas con la búsqueda contra la base de datos señuelo (invertida) y el número total de espectros, e interpola en esta distribución la puntuación obtenida con la búsqueda contra la base de datos objetivo, con objeto de determinar el *p*-value. Es importante puntualizar que, si estamos trabajando con un conjunto de datos de 100.000 espectros, el valor más bajo que se va a

poder estimar de la probabilidad será 1/100.000, o sea, 10^{-5}, que corresponderá al más alto de las mejores puntuaciones obtenidas de entre todos los espectros cuando se buscan contra la base de datos inversa. Por tanto, al buscar contra la base de datos normal, como se van a obtener puntuaciones más altas que la más alta de la base de datos inversa, en vez de extrapolar, lo que indicamos es que la probabilidad asociada a esa puntuación es menor que 10^{-5} en este caso.

En el caso de la implementación del método de la calidad única, el fichero que creamos lo llamamos "*quality file*". Este fichero almacena la información de las 1.000 primeras mejores puntuaciones para cada espectro. Este método determina las funciones *I(x,Q)* para cada espectro, por un ajuste por mínimos cuadrados a la curva formada por la puntuación tabulada contra la posición relativa en el ranking (normalizado entre 0 y 1). Para realizar este ajuste, se eliminó en todos los casos el primer punto, correspondiente a la mejor puntuación, ya que, al realizarse una única búsqueda contra una base de datos normal, la mejor puntuación en muchos casos corresponde a una secuencia correcta y por tanto no aleatoria. Luego, se calcula la probabilidad asociada a la mejor puntuación, esto es, el valor de $I_N(x,Q)$, usando la ecuación de escalado, siento *N* el número total de secuencias peptídicas candidatas para cada espectro. La interpolación en las distribuciones de cada espectro se realizó, en una primera etapa, de forma no paramétrica, y posteriormente, ajustando las distribuciones a la siguiente función, seleccionada por ajustarse bien a las distribuciones experimentales de las puntuaciones de varios espectros seleccionados al azar, y por ser además una función fácilmente derivable:

$$I(x,Q) = e^{\propto x^2+\beta x+\gamma}$$

Esta función, puede ser ajustada siempre que se tenga un número mínimo de puntuaciones, y luego se extrapola para conocer su valor para la mejor puntuación. Éste método no deja de ser una extrapolación, con los consiguientes posibles errores que se pueden cometer. Sin embargo, comparando los resultados obtenidos con los de otros métodos estadísticos empíricos, con colecciones de espectros de varios proteomas, parece que estima lo suficientemente bien las probabilidades de las asignaciones de cada espectro. En la Figura 8 se muestra, en un caso en concreto de un espectro, cómo se puede ajustar *I(x,Q)* con la fórmula de la exponencial, y se ve cómo de los 1.000 primeros scores, los primeros se desvían ligeramente, lo que da a entender que no siguen un comportamiento puramente aleatorio. Por ello no se toma en cuenta el primer punto para el ajuste. El tamaño de muestra (1.000 puntos en total) evita que la desviaciones de los espectros mejor puntuados influya en el resultado del ajuste.

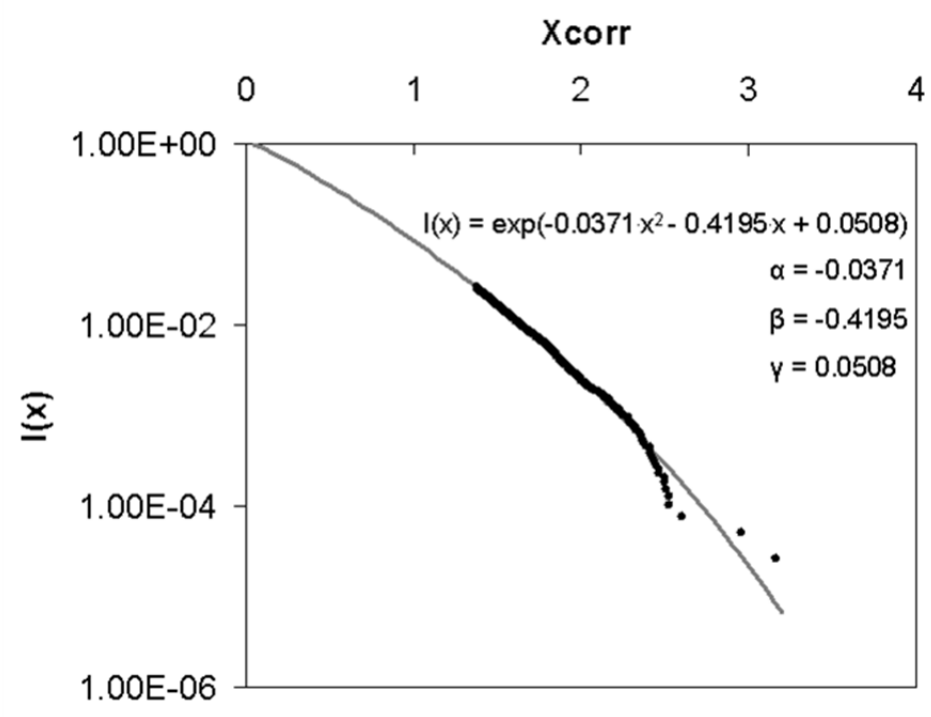


Figura 8. Ajuste de la función I(x,Q) de un espectro concreto. La figura muestra los parámetros de ajuste de la función I(x,Q) de un espectro en particular seleccionado al azar. Con puntos negros se muestran las 1.000 primeras mejores puntuaciones del espectro al buscarlo contra una base de datos real, y con la línea gris se representa la curva ajustada sobre esos puntos, exceptuando el primero de ellos.

Para evitar problemas de estimación de los parámetros en algunos casos, sobre todo en aquellos espectros con una baja calidad en los que SEQUEST no es capaz de encontrar 1.000 puntuaciones, se establecieron ciertas restricciones. Por un lado, si el valor de α es positivo, indicando que la curva es cóncava y no convexa según lo esperado (figura anterior), se repite el ajuste forzando α=0. Por otro lado, si alguno de los coeficientes α y β adquiere un valor menor a -10, se considera que el ajuste a esta función tampoco es válido y por tanto, se fuerza el valor de *I(x)*, y en consecuencia también de $I_N(x)$, a uno.

El método de la calidad única permite calcular la tasa de error o FDR de manera directa, a partir de los resultados obtenidos al buscar contra la base de datos real, usando la ecuación:

$$FDR_p = \frac{T - O_p}{O_p} \times \frac{p}{1 - p}$$

Donde $p \equiv I_N(x)$, T es el número total de espectros, y O_p es el número de espectros observados con una probabilidad igual o menor que p.

3.2. Materiales y métodos para el desarrollo de herramientas basadas en estándares HUPO-PSI

3.2.1. Experimento multi-centro 6 de ProteoRed (PME6)

Preparación de muestras y adquisición de datos por espectrometría de masas

Con el experimento multi-centro nº 6 de ProteoRed (PME6) se quisieron evaluar las capacidades y rendimientos de los diferentes flujos de trabajo y tecnologías aplicadas por cada uno de los participantes a la hora de analizar una mezcla de complejidad media, con 4 proteínas añadidas a diferentes concentraciones y un fondo de proteínas provenientes del plasma humano, con distintas isoformas. La muestra, llamada ASS17v3 o PPSR (*ProteoRed Plasma Subset Reference*) (Figura 9), consiste en un conjunto (*pool*) de muestras de plasma humano provenientes de sujetos sanos al que primeramente se le aplicó una inmunodeplección usando la columna SEPPRO IgY4 (Sigma) para, en este caso, capturar las 14 proteínas más abundantes. Posteriormente, las fracciones resultantes de 9 de esas depleciones se combinaron, se pasaron por un filtro de centrífuga AMICON (Millipore) y fueron inmunodepleccionadas de nuevo con la columna SEPPRO IgYHSA (Sigma) para eliminar la mayoría de la proteína Albúmina (*HSA, Human Serum Albumina*), la más abundante de las proteínas de plasma. Finalmente, a un total de 2.967 μg se le añadieron 30 μg de YWHAG (*P61981|1433_HUMAN*: Proteína 14-3-3 gamma, recombinante de humano), 3 μg de ALDOA (*P00883|ALDOA_RABIT*: Aldolasa fructosa bifosfato A de conejo, Sigma), 0,3 μg de CASB (*P02666|CASB_BOVIN*: Beta caseína bovina, Sigma) y 0,03 μg de PYGM (*P00489|PYGM_RABIT*: Glicógeno fosforilasa de conejo, Sigma), es decir, 4 proteínas con 5 órdenes de magnitud de diferencia en concentración. Esta muestra, consistente en una mezcla de alrededor de 145 proteínas de plasma humano más las 4 proteínas añadidas fue repartida a 20 participantes. Cada participante suspendió la muestra en 50mM Tris HCl pH 7.4 y la almacenó a -20ºC. Cada participante realizó su propio gradiente cromatográfico, pero se recomendó la utilización de un gradiente de entre 90 y 120 minutos para obtener datos lo más comparables posibles.

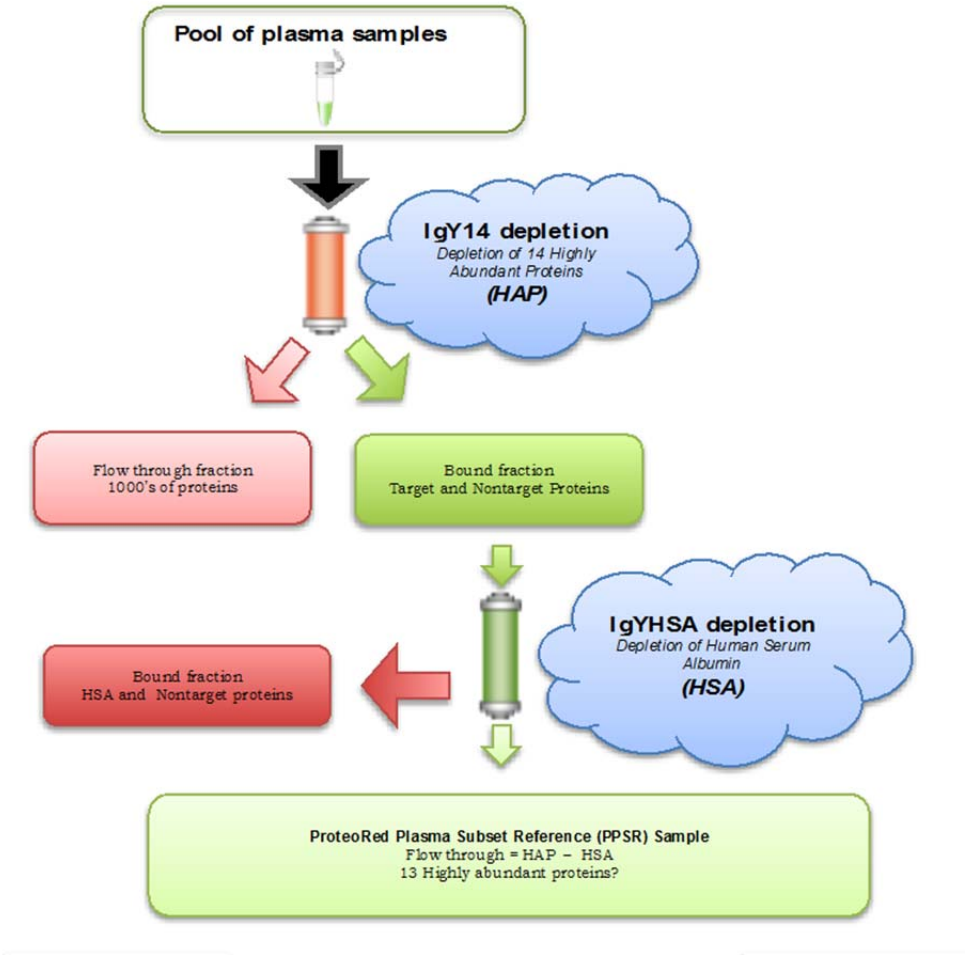


Figura 9. Flujo de la preparación de la muestra PPSR (ProteoRed Plasma Subset Reference): un conjunto de muestras de plasma humano al que primeramente se le aplicó una inmunodeplección usando la columna SEPPRO IgY4 (Sigma) para, capturar las 14 proteínas más abundantes. Posteriormente, 9 fracciones se combinaron, se pasaron por un filtro AMICON (Millipore) y se inmunodepleccionaron de nuevo con la columna SEPPRO IgYHSA (Sigma) para eliminar la mayoría de la proteína Albúmina. Finalmente, a un total de 2967 µg se le añadieron 4 proteínas a diferentes concentraciones: 30 µg de YWHAG, 3 µg de ALDOA, 0,3 µg de CASB y 0,03 µg de PYGM.

Cada participante empleó también su propio sistema de espectrometría de masas según se muestra en la tabla (Tabla 5):

#	Institution CODE	Orbitrap	Triple quadrupole	Q-TOF	MALDI-TOF TOF	Ion Trap	Linear Ion Trap
1	CNB				4800 MALDI-TOFTOF (ABSciex)	HCT Ultra 3D ion-trap (Bruker)	
2	CBT			Qstar Elite (ESI-Q-TOF) (ABSciex)			
3	LP-CSIC/UAB	LTQ-Orbitrap (Thermo)					
4	UPF	LTQ-Orbitrap-Velos (Thermo)					
5	CBM						LTQ Velos (Thermo)
6	HUVH	LTQ-Orbitrap (Thermo)				Ion Trap Esquire-Ultra (Bruker)	
7	UPV-EHU			Synapt HD (Waters)			
8	UA					Ion Trap XCT plus (Agilent)	
9	PCM-UCM				4800 MALDI-TOFTOF (ABSciex)		LTQ (Thermo)
10	UCO	LTQ-Orbitrap (Thermo)	4000 Qtrap (ABSciex)				
11	CIPF			QSTAR XL (ABSciex)			
12	UB	LTQ-Orbitrap-Velos (Thermo)					
13	CIMA			QTOF Micro (Waters)			
14	I+CS						
15	CMU	LTQ-Orbitrap-Velos (Thermo)			4800 MALDI-TOFTOF (ABSciex)		
16	CIB	LTQ-Orbitrap-Velos (Thermo)					
17	INIBIC				4800 MALDI-TOFTOF (ABSciex)		
18	CIC					LCQ DecaXP	
19	PCB	LTQ-Orbitrap-Velos (Thermo)					
20	CIC bioGUNE			QToF Premier (Waters)			

Tabla 5. Equipamiento de los participantes en el experimento multi-centro 6 de ProteoRed (PME6): Un total de 20 laboratorios, dos de ellos extranjeros, participaron en el experimento multi-centro 6 de ProteoRed utilizando diferentes plataformas de espectrometría de masas de diferentes casas comerciales.

Los resultados obtenidos por cada laboratorio fueron en su día recopilados usando la herramienta online generadora de documentos MIAPE (sección 4.2.1), dentro del proyecto MIAPE público con identificador "470". Sin embargo, dichos MIAPE no contenían

explícitamente los resultados de identificación (algunos contenían un enlace a una tabla de resultados), ya que la automatización aún no era posible sin la herramienta MIAPE Extractor (descrita en la sección 4.2.5). Para hacer la comparativa de resultados se recopilaron los datos por medio de una plantilla Excel creada por el grupo de trabajo organizador del experimento, disponibles aquí: http://www.proteored.org/PME6_Results.asp.

Reprocesado de datos

Por un lado, a partir de las platillas Excel enviadas por cada participante se crearon unos ficheros de texto separados por comas que contenían únicamente la información proteína-péptido-puntuación de péptido. Dichos ficheros se utilizaron para introducir en el MIAPE Extractor la información enviada por cada participante.

Adicionalmente se realizó un reanálisis centralizado para obtener datos comparativos más robustos a partir de los datos crudos de cada participante. Para ello, los fichero binarios de datos crudos fueron transformados, dependiendo del caso, o bien en formato MGF (*Mascot Generic Format*), o bien en el estándar PSI para datos crudos, el mzML versión 1.1, utilizando para ello la herramienta msconvert de ProteoWizard (Kessner, Chambers et al. 2008) versiones 3.0.4360 y 3.0.4624 – 32bits y la herramienta CompassXport versión 3.0.5 de Bruker. Dichos datos fueron analizados con el motor de búsqueda Mascot (*Matrix Science*) en sus versiones 2.3 y 2.4. La base de datos utilizada en el caso del experimento multi-centro 6 de ProteoRed fue la base de datos Uniprot-Swissprot con taxonomía humana a la cual se le añadieron las secuencias de las proteínas añadidas a la muestra (http://www.proteored.org/PME6_Databases.asp). Se utilizó una base de datos señuelo concatenada con la base de datos normal, invirtiendo los péptidos trípticos de la base de datos objetivo para formar el señuelo. Los resultados de Mascot fueron exportados por medio de un script escrito en Perl, "*export2mzid11.pl*" (*Matrix Science*) a ficheros estándares PSI mzIdentML versión 1.0 y 1.1 (desde Mascot 2.3 y 2.4 respectivamente).

3.2.2. Programación y hardware

La implementación de las herramientas web se realizó con el lenguaje de programación ASP (*Active Server Pages*) y Javascript, probándose en diferentes navegadores como Microsoft Internet Explorer, Firefox, Safari o Chrome.

La base de datos de documentos MIAPE está diseñada e implementada sobre una base de datos SQL Server 2008.

El desarrollo de la API (*Application Programming Interface*) para manejar información MIAPE se basa en el lenguaje Java 6.0 y utiliza distintas tecnologías Java: para el manejo de XMLs, JAXB 2.1, y para el acceso a bases de datos, Hibernate 3.0.

Los servicios web desplegados también fueron programados en Java, utilizando la API de MIAPEs y sirviendo por medio de la tecnología SOAP (Axis 1.4). Dos ficheros de descripción de los servicios web, WSDL están disponibles públicamente para posibilitar la conexión de clientes externos (http://proteo.cnb.csic.es:9999/miape-api-webservice/MiapeAPIWebservicePort?WSDL y http://proteo.cnb.csic.es:9999/miape-extractor-webservice/MiapeExtractorPort?WSDL).

El desarrollo Java se realizó en un entorno de programación MyEclipse Java Enterprise SDK versión Indigo 3.7.2.

Para el manejo programático de ficheros estándares mzML y mzIdentML se utilizaron las librerías escritas en Java jmzML (Cote, Reisinger et al. 2010) y jmzIdentML (Reisinger, Krishna et al. 2012), respectivamente. Para el manejo de los ficheros XML de resultados del motor de búsqueda X!Tandem se utilizó la librería en Java XTandem-Parser (Muth, Vaudel et al. 2010).

3.2.3. Métodos de análisis en el MIAPE Extractor

La herramienta MIAPE Extractor extrae la información MIAPE, en este caso, de identificación de péptidos y proteínas, de ficheros estándares y la almacena en una base de datos de documentos MIAPE. Adicionalmente, la herramienta permite agregar, inspeccionar y analizar los datos de diferentes búsquedas a un mismo tiempo, pudiéndose comparar datos de diferentes experimentos o de diferentes búsquedas realizadas sobre diferentes fracciones o bandas en un experimento con pre-fraccionamiento. El usuario decide qué datos quiere inspeccionar o analizar, y para ello define un proyecto de análisis en forma de árbol, organizando el conjunto de MIAPEs MSI, cada uno equivalente a un resultado de un motor de búsqueda, es decir, una lista de péptidos y proteínas, en tres niveles (Figura 10):

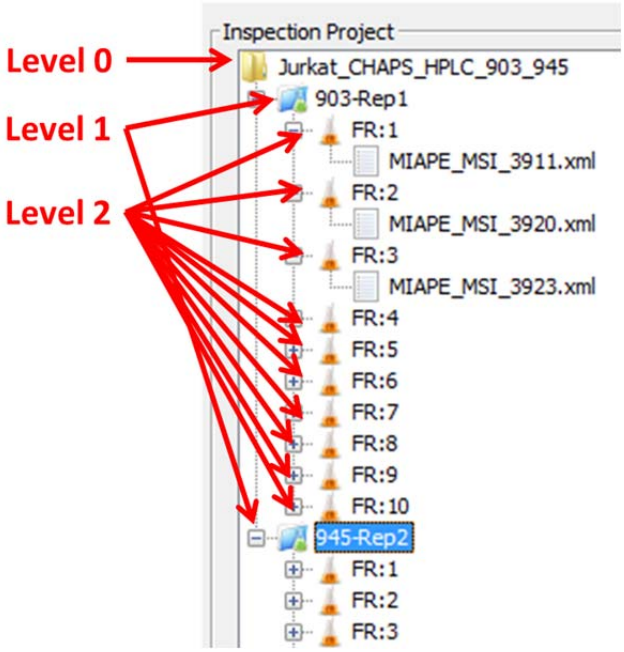


Figura 10. Ejemplo de organización de datos analizados en el MIAPE Extractor: el nivel 0 agrupa dos experimentos réplicas biológicas del análisis de una misma muestra. El nivel 1 agrupa cada una de las dos réplicas, conteniendo 10 fracciones cada una. El nivel 2, corresponde a cada una de las búsquedas realizadas con los espectros adquiridos para cada una de las fracciones.

- Nivel 2: en este nivel se asocia uno o varios MIAPEs MSI, es decir, uno o varios resultados de un motor de búsqueda. Normalmente, este nivel corresponde al nivel de fracciones o bandas en un experimento con pre-fraccionamiento.
- Nivel 1: en este nivel se agrupan varios nodos del nivel 2. Normalmente este nivel corresponde a un experimento que contiene diferentes búsquedas provenientes de un pre-fraccionamiento.
- Nivel 0: este nivel agrupa una colección de nodos de nivel 1 y normalmente corresponde a una colección de experimentos con interés de agregar todos sus datos a un mismo nivel. Pueden ser réplicas técnicas o biológicas, o bien, simplemente experimentos diferentes sobre una misma muestra.

Agregado de datos

Como se ha descrito, los datos introducidos por la herramienta MIAPE Extractor se pueden agrupar y analizar a distintos niveles. Cada MIAPE MSI analizado corresponde a una lista identificada de péptidos y proteínas. Los MIAPEs MSI en un mismo nodo de nivel 2 son considerados como una única lista de péptidos y proteínas. Los datos mostrados en cada nivel

superior, esto es, niveles 1 y 0, son resultado de unir las listas de identificaciones de los niveles inferiores (Figura 11).

A partir de la versión 3.1.0 de la herramienta MIAPE Extractor, la inferencia de proteínas es revisada por el programa de tal manera que se ejecuta el algoritmo de agrupamiento de proteínas PAnalyzer (Prieto, Aloria et al. 2012). Este algoritmo agrupa las proteínas que comparten algún péptido, y clasifica luego los grupos obtenidos en diferentes tipos: grupos conclusivos, grupos no conclusivos, grupos ambiguos, y grupos indistinguibles (Figura 11).

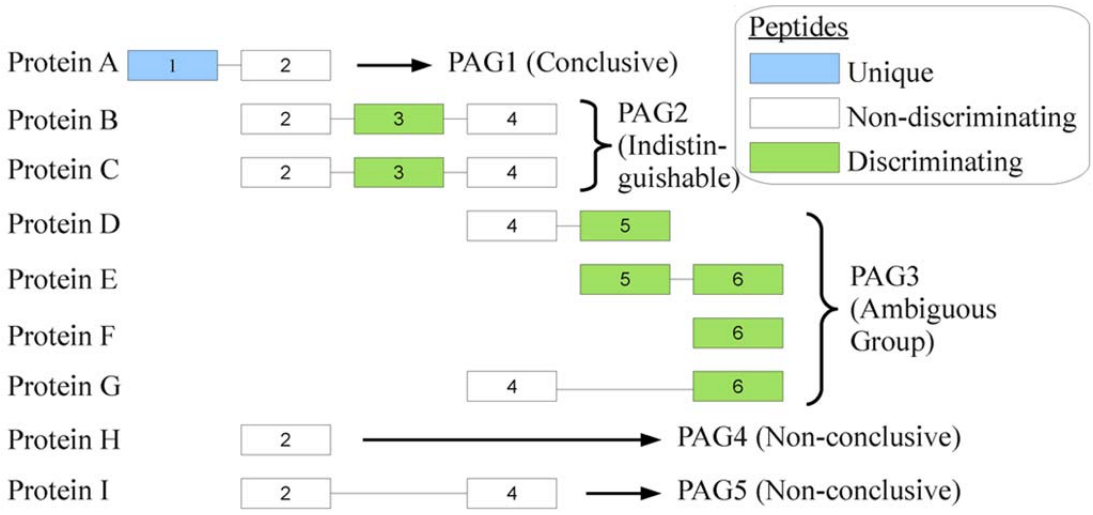


Figura 11. Clasificación de los agrupamientos según el algoritmo PAnalyzer.

El reagrupamiento de las proteínas se realiza siempre en cada uno de los niveles que estamos analizando, salvo en el nivel 0, en cuyo caso es opcional realizar el agrupamiento de todos los datos de niveles 1 y 2.

Cálculo de la tasa de error FDR

La herramienta MIAPE Extractor es capaz de aplicar diferentes filtros a las listas de péptidos y proteínas analizadas, entre ellos, un filtro por FDR o tasa de error. Este filtro se realiza en el nivel 2, es decir, aplicado a cada resultado de un motor de búsqueda individualmente, ya que la tasa de error debe ser entendida como la tasa de error en cada una de las búsquedas realizadas y no como la tasa de error de los resultados agregados de varias búsquedas. Así pues, la lista de péptidos y proteínas resultante en el nivel 1 cuando se aplica un filtro por FDR en el nivel 2, será la lista resultante de combinar las listas de péptidos y proteínas resultantes del filtro por FDR a nivel 2 (Figura 12).

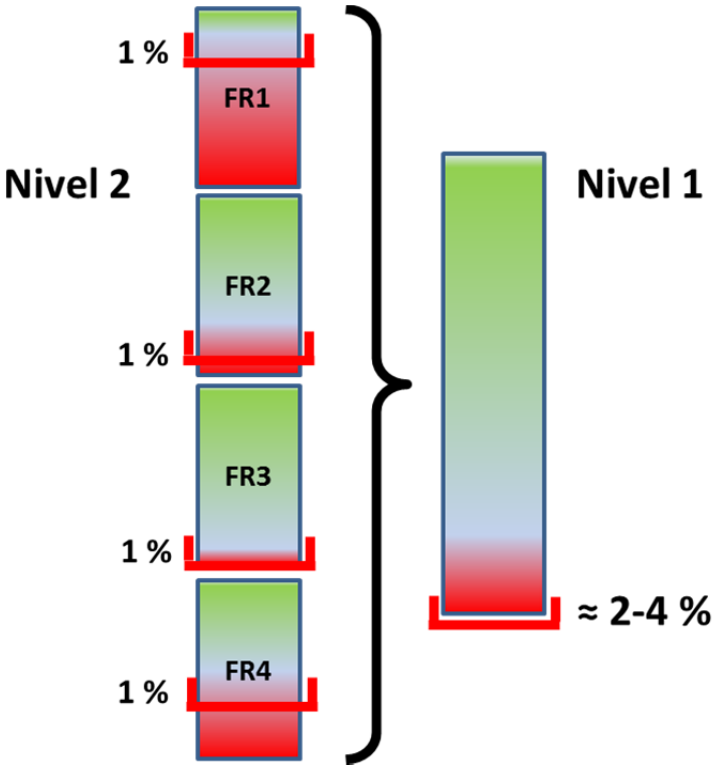


Figura 12. Ejemplo de cálculo de la FDR y su agregación por niveles en el MIAPE Extractor: el corte por FDR de 1% se realiza en el nivel 2. Los datos resultantes de la aplicación del filtro en cada nodo del nivel 2 se pasan al nivel 1. En el nivel 1, se calcula de nuevo la tasa de error, que normalmente suele ser superior al filtro aplicado en el nivel inferior, entre un 2% o un 4 %.

El filtro por FDR aplicado por la herramienta puede ser a nivel de PSM, de péptido o de proteína. A nivel de PSM, todos los PSMs se ordenan según la puntuación seleccionada y se calcula la tasa de error. A nivel de péptido, se agrupan todos los PSMs asignados a una misma secuencia y se ordenan teniendo en cuenta únicamente el mejor de ellos para calcular la FDR. En el caso de la FDR a nivel de proteína, primeramente se optó por una aproximación simplificada, consistente en agrupar los PSMs pertenecientes a una misma proteína (o grupo) y teniendo en cuenta únicamente el de mejor puntuación de cada una (o cada uno) para hacer el ordenamiento de proteínas. Pese a que éste método puede ser criticable por diversas razones (por ejemplo, no tiene en cuenta el hecho de que una proteína identificada con dos péptidos es más probable que una identificada con sólo uno pero de mejor puntuación), se optó por él en primera instancia por su gran sencillez. En el momento de escribir esta tesis se está empezando a implementar otro método, basado en un ordenamiento de las proteínas según una puntuación equivalente al producto de la puntuación PEP de los péptidos asociados a cada proteína, siendo estos valores PEP a nivel de péptido calculados a partir de los valores PEP a nivel de PSMs, y éstos calculados a partir de las puntuaciones *E*-value.

Resultados

4. Resultados

4.1. Desarrollo de métodos de validación de identificaciones de péptidos y proteínas a gran escala por espectrometría de masas

Como se ha comentado en la introducción, la Proteómica moderna o de segunda generación se basa en el análisis de proteínas por espectrometría de masas y, gracias al uso cada vez más común de técnicas de separación multi-dimensionales acopladas a la espectrometría, el número de identificaciones de péptidos y proteínas en un mismo experimento es cada vez mayor y se hace cada vez más necesario un modelo de puntuación para discernir, de la manera más automática posible, las identificaciones correctas de las que son falsos positivos.

En este trabajo nos referiremos a las distribuciones de las mejores puntuaciones construidas a partir de una gran colección de espectros como las distribuciones promedio de puntuaciones (*average score distributions*), mientras que las distribuciones de puntuaciones asociadas a un único espectro las llamaremos distribuciones individuales de puntuaciones (*single-spectrum score distributions*) (Figura 13). A pesar de que las distribuciones promedio de puntuaciones han sido utilizadas muy frecuentemente, algunas propiedades y comportamientos como su dependencia del tamaño de la base de datos utilizada y en concreto las particularmente relacionadas con las puntuaciones de SEQUEST no se han analizado ni entendido anteriormente a este estudio. Por ejemplo, no existe un criterio claro sobre cómo utilizar la mejor puntuación *XCorr* y la puntuación ΔC_n. De hecho, la puntuación ΔC_n ha sido utilizada en varios trabajos como criterio de corte de tal manera que sólo las asignaciones con una puntuación delta determinada mínima son consideradas como positivos potenciales (Link, Eng et al. 1999, Washburn, Wolters et al. 2001, Florens, Washburn et al. 2002, Peng, Elias et al. 2003, Qian, Liu et al. 2005), y en otros trabajos se ha utilizado como un parámetro adicional e independiente de la puntuación mejor (Keller, Nesvizhskii et al. 2002, Kislinger, Rahman et al. 2003, Lopez-Ferrer, Martinez-Bartolome et al. 2004). Pese a que existe una obvia relación entre la mejor puntuación y la puntuación delta (que es la diferencia entre la mejor puntuación y la segunda mejor), no se ha analizado aún la relación teórica existente entre estos dos parámetros. De la misma manera, tampoco se ha descrito la relación entre las distribuciones promedio y las distribuciones individuales cuando se usa el mismo esquema de puntuaciones.

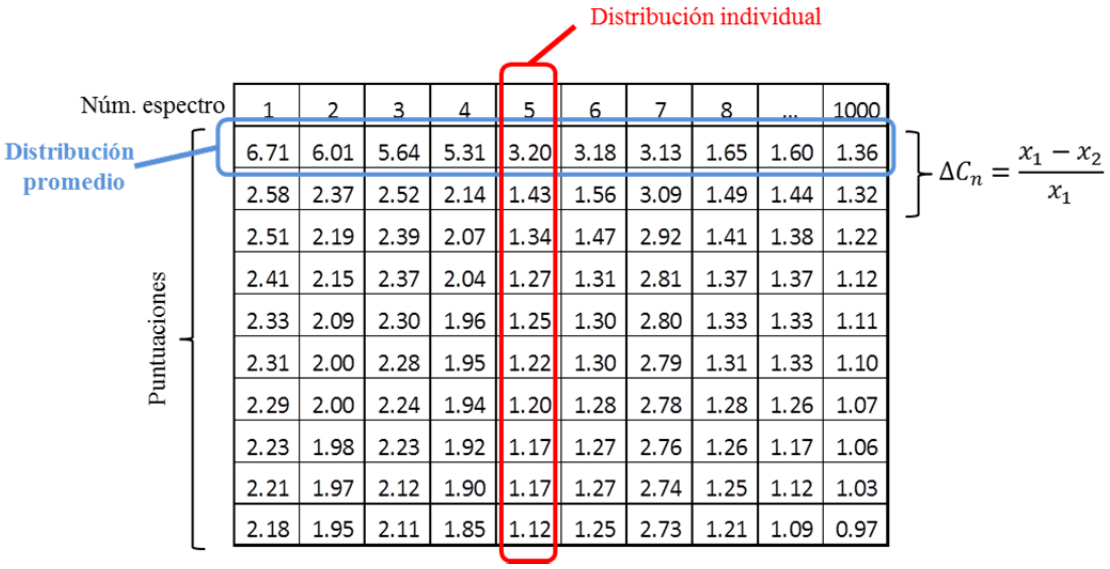


Figura 13. Esquema de creación de las distribuciones promedio de puntuaciones e individuales de puntuaciones: La distribución promedio de puntuaciones (azul) está formada por la mejor puntuación obtenida por cada uno de los, en este caso, 1.000 espectros. La distribución individual de puntuaciones (rojo) está formada por las puntuaciones de las mejores asignaciones a cada espectro individual.

4.1.1. La ecuación de escalado

Para cada uno de los espectros de masas, definimos la distribución de probabilidad *I(x)* como una función que describe la probabilidad de obtener una puntuación igual o mejor que x, cuando un espectro se busca contra una secuencia peptídica seleccionada al azar. Por otro lado, la mayor puntuación obtenida al buscar un espectro contra un conjunto de secuencias al azar, es decir, la mejor asignación de todas para un espectro concreto, corresponde a la mejor puntuación obtenida cuando la búsqueda se repite *N* veces, siendo *N* el número de secuencias candidatas. La probabilidad de que dado un espectro se obtenga una puntuación igual o mejor que x cuando se busca contra *N* secuencias, esto es, $I_N(x)$, está relacionada con la probabilidad de obtener esa misma puntuación contra únicamente una secuencia, esto es, *I(x)*, mediante lo que denominaremos la ecuación de escalado:

$$I_N(x) = 1 - (1 - I(x))^N$$

En general, esta ecuación es válida para la j-ésima mejor puntuación,

$$K_N^{(j)}(x) = 1 - \sum_{i=0}^{i=j-1} \binom{N}{i} I(x)^i (1 - I(x))^{N-i}$$

siendo $K_N^j(x)$ la probabilidad de obtener la j-ésima mejor puntuación al buscar el espectro contra *N* secuencias candidatas. Además, se espera que cada espectro tenga una distribución

individual diferente al resto de espectros, teniendo un comportamiento diferente cuando se compara con las secuencias candidatas.

Para comprobar la ecuación de escalado, seleccionamos 6 espectros con diferentes puntuaciones XCorr (al buscar con SEQUEST) de un proteoma de núcleos de células Jurkat (Lopez-Ferrer, Martinez-Bartolome et al. 2004). Generamos 10.000 secuencias peptídicas aleatorias por Monte Carlo y buscamos esos 6 espectros contra dichas secuencias utilizando el buscador SEQUEST y construimos las funciones de probabilidad *I(x)*. Luego, agrupamos las 10.000 secuencias candidatas en 10 grupos de 1.000 secuencias y por otro lado, en 100 grupos de 100 secuencias. Con la primera agrupación construimos la función de probabilidad para la mejor puntuación de cada uno de los 10 grupos $I_{1000}(x)$. Con la segunda, construimos la función de probabilidad para la quinta mejor puntuación de cada uno de los 100 grupos, que llamaremos $K_{100}^{5}(x)$. Como se muestra en la Figura 14, aplicando la ecuación de escalado a *I(x)* obtenemos en todos los casos una predicción perfecta de las distribuciones experimentales. Igualmente, se obtienen los mismos resultados analizando las distribuciones de las puntuaciones obtenidas en otras posiciones en el ranking contra diferentes números de secuencias candidatas (no se muestran). Estos resultados demuestran la validez de la ecuación de escalado.

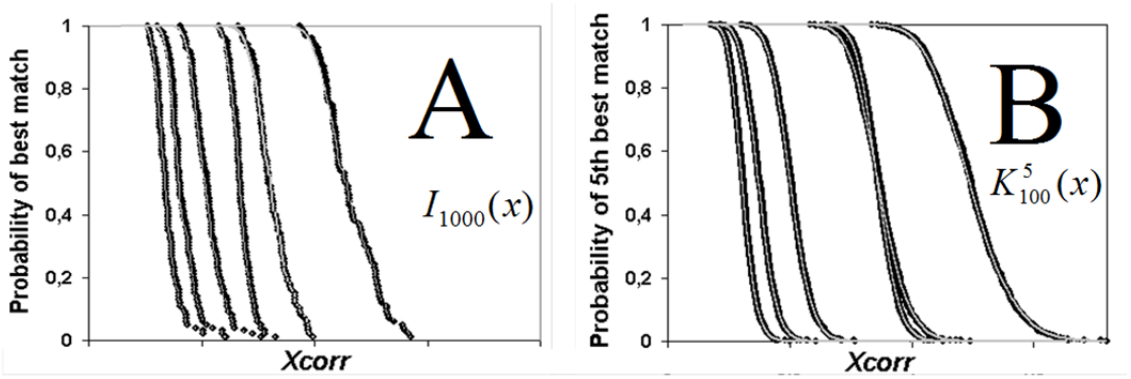


Figura 14. Distribución de probabilidad y la ecuación de escalado: Cada uno de los 6 espectros seleccionados al azar se buscaron contra 10.000 secuencias aleatorias generadas por Monte Carlo, y se analizaron las puntuaciones de SEQUEST XCorr. (A-B) Distribuciones de probabilidad obtenidas a partir de la mejor y de la quinta mejor puntuación cuando éstos se agrupan en conjuntos de N=1.000 (A) o N=100 (B), respectivamente (puntos negros). Las líneas grises representan las distribuciones estimadas calculadas para la distribución con N=1 y aplicando la ecuación de escalado.

4.1.2. Distribuciones promedio de probabilidad y calidad del espectro

A las distribuciones promedio de probabilidad, formadas por las mejores puntuaciones derivadas de buscar un conjunto grande de espectros contra una base de datos, las

denominaremos $\boldsymbol{I_N(x)}$ en negrita, donde *N* es otra vez el número de secuencias contra las cuales se buscan los espectros, y es un parámetro que, obviamente, depende del tamaño de la base de datos contra la que se busque.

Esta aproximación asume implícitamente que todos los espectros tienen un comportamiento similar de asignación aleatoria. Como se muestra en la Figura 15, esta premisa es incorrecta, al menos aplicada a las puntaciones de SEQUEST, y podría no considerarse una aproximación válida a una situación real. Esta figura nos permite ver que las distribuciones de probabilidad individuales de la mejor puntuación de SEQUEST (XCorr) determinadas experimentalmente para los seis espectros de masas escogidos (líneas grises) tienen una pendiente más pronunciada que la distribución promedio de probabilidad determinada a partir de todos los espectros del proteoma (líneas negras gruesas), y los puntos de inflexión de las curvas (que denominaremos las puntuaciones de transición) de cada espectro están claramente posicionados en diferentes valores de XCorr (Figura 15 A). De la misma manera, las correspondientes distribuciones de densidad de probabilidad de cada espectro son mucho más estrechas que las distribuciones promedio (Figura 15 B), sugiriéndonos que los valores más probables para las mejores puntuaciones, para cada espectro en particular, están limitados en unos rangos de puntuación bastante estrechos y pueden tomar valores muy diferentes de un espectro a otro.

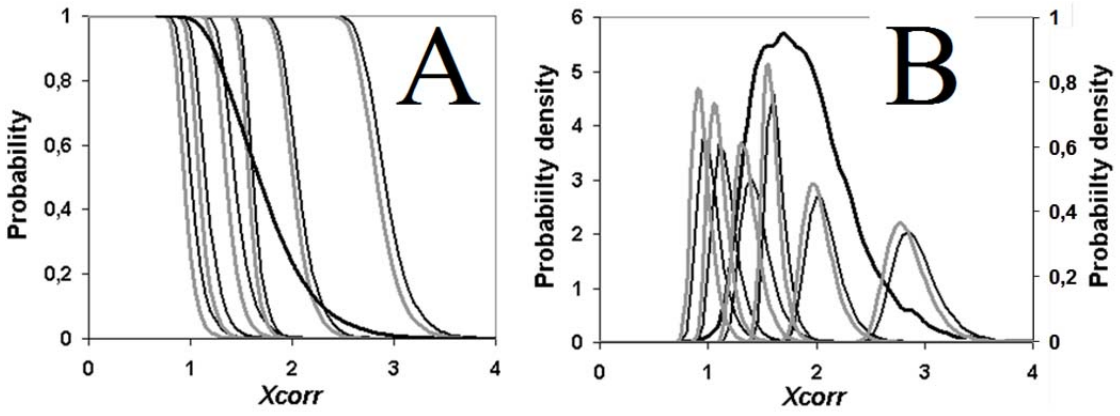


Figura 15. Distribución de probabilidad y la ecuación de escalado: Cada uno de los 6 espectros seleccionados al azar se buscaron contra 10.000 secuencias aleatorias generadas por Monte Carlo, y se analizaron las puntuaciones de SEQUEST Xcorr. (A) Distribuciones de probabilidad predichas y (B) distribuciones de densidad de probabilidad predichas de la mejor puntuación para cada uno de los 6 espectros obtenidas después de buscar cada uno de ellos contra el mismo número de secuencias candidatas aleatorias contenidas en la base de datos human.fasta (líneas grises), o usando el método de calidad única (líneas negras). Superponiendo la distribución de probabilidad promedio y la distribución de densidad de probabilidad de la mejor puntuación obtenida después de analizar los 40.000 espectros del proteoma modelo (líneas en negrita); en (B) las distribuciones están dibujadas a diferentes escalas, como muestra el eje de ordenadas derecho.

Los espectros se clasifican por tanto en una serie de categorías determinadas por un parámetro que llamaremos *calidad, Q*. Esto se corresponde con la noción intuitiva de que un espectro con una calidad mayor tenderá a tener mejores puntuaciones por azar. Esta parametrización clasifica los espectros en subconjuntos de espectros que comparten la misma distribución de probabilidad.

Así pues, la distribución de probabilidad de una colección de espectros, esto es, la distribución promedio de las puntuaciones, se puede expresar como una combinación de las distribuciones de probabilidad de las mejores puntuaciones de los espectros que la componen, de acuerdo con la distribución de calidades de la muestra:

$$I(x) = \sum_{Q=1}^{q} f(Q) I(x, Q)$$

La definición de calidad asegura que la distribución promedio de probabilidad de la puntuación de un conjunto de espectros dentro de un mismo subconjunto de calidad, cumplirá las ecuaciones de escalado. Sin embargo, las ecuaciones de escalado, en general, no pueden aplicarse a las distribuciones promedio de probabilidad.

Región de posible identificación

Definimos la región de posible identificación como la región del espacio de puntuaciones donde se cumple que:

$$N \times \boldsymbol{I}(\boldsymbol{x}) < 1$$

Con esta definición, delimitamos una región que contiene al menos todas las puntuaciones cuyas asignaciones de péptidos pueden ser consideradas como estadísticamente significativas.

En un trabajo teórico realizado por el Dr. Fernando Martín Maroto, se demostró que, bajo condiciones generales, cuando una puntuación cae dentro de la región de posible identificación, las distribuciones promedio de probabilidad verifican la ecuación de escalado. En la práctica, esto significa que en la región de baja probabilidad, podemos utilizar las ecuaciones de escalado para recalcular la probabilidad de una puntuación dada, cuando el tamaño de la base de datos cambia.

Más adelante se describirá un método, el método de la calidad única (ver apartado 4.1.4), derivado de la aplicación de la ecuación de escalado. Este método se basa en la construcción de la distribución de probabilidad individual de cada espectro y la aplicación de la ecuación de escalado considerando el número *N* total de secuencias peptídicas candidatas contra las cuales

se buscan cada uno de los espectros. Luego, estas distribuciones se utilizan para calcular la significatividad estadística de la mejor puntuación de cada espectro.

La relación de escalado puede usarse para extrapolar desde una distribución $\boldsymbol{I_A(x)}$ a otra $\boldsymbol{I_B(x)}$ si el tamaño de la base de datos B es menor que A o si B es mayor, pero no mucho mayor que A, ya que, en este caso, $A \times \boldsymbol{I(x)} < 1$ puede que no implique necesariamente que $B \times \boldsymbol{I(x)} < 1$. En general, y puesto que *N* es proporcional al tamaño de la base de datos, la siguiente ecuación se puede usar para estimar la probabilidad de la mejor puntuación en una base de datos de tamaño *S* siempre que se conozca la distribución promedio de las puntuaciones para una base de datos de *test* de tamaño S_{TEST}:

$$\mathbf{I}_S(x) = 1 - (1 - \mathbf{I}_{S_{TEST}}(x))^{\frac{S}{S_{TEST}}}$$

En (Figura 16) se muestra la validez de esta ecuación en la región de posible identificación. Se puede ver que la cola de las distribuciones promedio de probabilidad de las mejores puntuaciones, obtenidas después de buscar los espectros contra bases de datos señuelo de diferentes tamaños, se puede predecir con precisión a partir de la distribución obtenida buscando contra la base de datos humana.

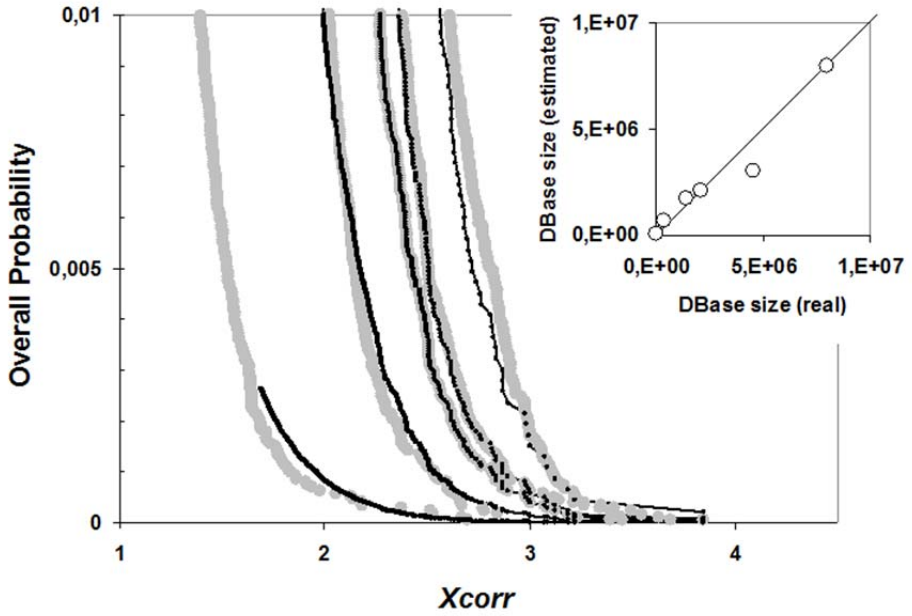


Figura 16. Validez de la ecuación de escalado para la distribución promedio de probabilidad en la región de baja probabilidad. Se muestran las distribuciones de probabilidad de las mejores puntuaciones obtenidas al buscar 40.000 espectros del proteoma modelo contra varias bases de datos como caballo, levadura, humano, swissprot y nr, respectivamente, de izquierda a derecha, y por orden de tamaño (puntos grises grandes). Los puntos negros pequeños representan las distribuciones calculadas aplicando la ecuación anterior a la distribución obtenida al buscar contra la base de datos humana, usando el tamaño de la base de datos que indica la mejor asignación en cada caso. En la gráfica interior, los tamaños de las bases de datos usados para representar las distribuciones estimadas, expresados como el número total de péptidos únicos, aparecen frente los tamaños reales de las bases de datos.

Fracción de calidad de una puntuación

Los espectros suelen buscarse contra bases de datos que suelen contener un gran número de secuencias candidatas; valores típicos para N son 40.000 para la base de datos humana y 200.000 para la base de datos nr. Como consecuencia de que N es un número grande, las distribuciones de probabilidad de las mejores puntuaciones para cada espectro individual se espera que lleguen a ser funciones de pendiente muy pronunciada, con un punto de inflexión, o puntuación de transición, donde la probabilidad sufre una brusca bajada, en claro contraste con el comportamiento de la probabilidad promedio (Figura 15 A). Las distribuciones de densidad de probabilidad para cada espectro se convierten en funciones muy estrechas, donde la puntuación de transición corresponde al valor más probable como mejor puntuación (Figura 15 B). Por tanto, la puntuación de transición se puede utilizar como un indicador de calidad, ya que los espectros que tengan mayor calidad tendrán el punto de inflexión localizado en puntuaciones más altas.

Otra consecuencia de que exista un gran número de secuencias candidatas es que dada una puntuación, es posible estimar cuál es la fracción esperada de espectros en la muestra que obtienen una puntuación igual o mejor de forma aleatoria; y nos referimos a esta fracción como la *fracción de calidad de una puntuación*. En otras palabras, el valor de la distribución promedio de probabilidad para una puntuación dada, es igual a la fracción de calidad de esa misma puntuación, reflejando la composición de calidades de una muestra. Este resultado implica que la distribución promedio de probabilidad de una colección de espectros dependerá únicamente de la distribución de calidades de los espectros.

Por el contrario, en la cola, dentro de la región de posible identificación, donde los valores de las puntuaciones no se espera que sean aleatorios y donde se satisface la ecuación de escalado, la distribución promedio de probabilidad depende principalmente de las distribuciones de puntuaciones de los espectros con mayor calidad. Es decir, los PSMs correctos siguen una distribución propia, separada de los PSMs incorrectos.

En la práctica, cuando un espectro se busca contra una base de datos real, el valor de la distribución promedio de probabilidad para la segunda mejor puntuación, puede considerarse como una estimación precisa de la fracción de espectros cuya calidad es mayor que la del espectro dado.

Esta idea se muestra en la Figura 17 A donde la distribución promedio de probabilidad se construye en combinación de las distribuciones de los seis espectros de masas. Luego, como se muestra en B, los valores dados por la distribución promedio de probabilidad para las segundas

mejores puntuaciones esperadas en cada uno de los espectros, reflejan de manera muy fehaciente la fracción de espectros que tienen mayor calidad.

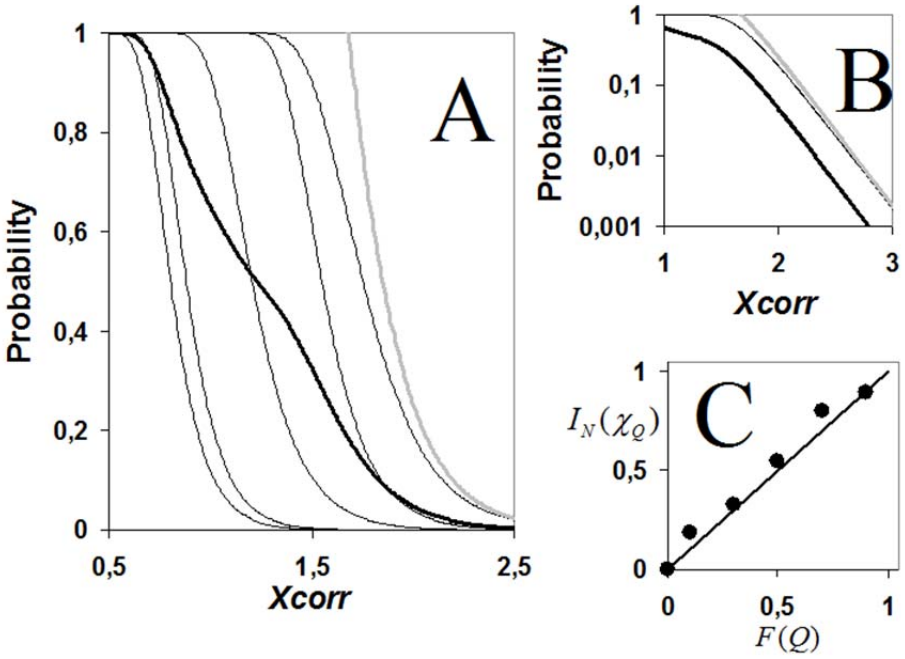


Figura 17. La probabilidad promedio de la segunda mejor puntuación refleja el factor de calidad de la puntuación. (A y B) La probabilidad promedio (línea negra gruesa) construida por la composición de las distribuciones de probabilidad de las mejores puntuaciones para cada uno de los seis espectros (líneas negras finas), asumiendo N=1.000. La línea gruesa en gris representa el cociente entre la probabilidad promedio y la probabilidad del valor más probable de la segunda mejor puntuación. Por simplicidad, en B sólo se muestra la distribución del espectro con el mayor componente de calidad. (C) Gráfica que muestra la probabilidad promedio de los valores más probables de las segundas mejores puntuaciones, para cada uno de los seis espectros, contra la fracción de calidad de la puntuación; la línea representa la identidad a lo largo de esas cantidades.

Teniendo en cuenta estos resultados de manera conjunta, se puede afirmar que no puede establecerse un criterio estadístico universal basado en determinados umbrales de corte sobre la mejor o la segunda mejor puntuación, porque la distribución de dichas puntuaciones siempre va a depender de la distribución de calidades, y por tanto, del conjunto de espectros con el que estemos trabajando.

4.1.3. Propiedades de las distribuciones promedio de puntuaciones de SEQUEST: el concepto de la razón de probabilidad

Como hemos dicho, vamos a considerar dos tipos de distribuciones de puntuaciones. La primera, característica de cada uno de los espectros MS/MS, es la distribución individual de

puntuaciones de espectro, la cual, para un valor de puntuación *x* obtenido al buscar un espectro contra *N* secuencias candidatas, nos da la probabilidad de que el espectro produzca una puntuación igual o mejor que *x* por azar. El segundo tipo de distribución, es la distribución promedio de puntuaciones $\mathbf{I_N(x)}$, que es característica de todos los espectros en el conjunto de datos que se considere y que se obtiene cuando una gran cantidad de espectros son buscados contra una base de datos señuelo. También vamos a considerar la distribución formada por las segundas mejores puntuaciones para cada espectro y la llamaremos $\mathbf{H_N(x)}$.

La distribución promedio de las mejores puntuaciones la consideramos, como hemos dicho, como la superposición de todas las distribuciones individuales de todos los espectros del conjunto de datos. Así pues, bajo este concepto, los espectros se clasifican de acuerdo con el parámetro que llamamos de *calidad*, cuya interpretación en la práctica es que los espectros con una calidad mayor tienden a dar mayores mejores puntuaciones tan sólo por azar. De aquí se derivan dos propiedades importantes:

- Los valores que toma la distribución promedio de puntuaciones $\mathbf{I_N(x)}$ tienden a ser proporcionales al número de secuencias candidatas cuando las probabilidades son suficientemente bajas, esto es, cuando las mejores puntuaciones toman los mayores valores y se espera que las asignaciones sean positivas.
- El valor tomado por la distribución promedio para la mejor puntuación de un espectro concreto $\mathbf{I_N(x_F)}$, es una estimación bastante precisa de la fracción de espectros en la colección total de espectros que tienen una calidad mejor o igual que ese espectro.

La distribución promedio de las segundas mejores puntuaciones $\mathbf{I_N(x_S)}$ tiene las mismas propiedades, y ya que la segunda mejor puntuación obtenida por un espectro cuando se busca contra una base de datos normal (no señuelo) es muy probable que sea una asignación por azar, el valor que toma la distribución promedio de las segundas mejores puntuaciones para la segunda mejor puntuación de un espectro puede ser utilizada como un estimador bastante preciso de la posición relativa que ocupa en la colección de espectros de acuerdo a su calidad.

Analizando cómo debería ser tratada la información de la mejor y de la segunda mejor puntuación para hacer inferencias estadísticas acerca de identificaciones a gran escala de péptidos, llegamos a la conclusión de que la probabilidad condicional de obtener una primera puntuación igual o mejor que x_F cuando se obtiene una segunda mejor puntuación igual o mejor que x_S se estima por la razón de la probabilidad promedio de la primera puntuación y la probabilidad promedio de la segunda puntuación, esto es, $\mathbf{I_N(x_F)/I_N(x_S)}$, la **razón de probabilidad (*PRatio*)**. Esta probabilidad condicional puede ser estimada también usando en el denominador la distribución de las segundas mejores puntuaciones, es decir, usando $\mathbf{H_N(x_S)}$.

Consideremos el siguiente ejemplo: si el 10% de los espectros de una colección es de alta calidad, y dada una puntuación, aparece al azar con una frecuencia de 0,2 (1 de cada 5 veces) en este subconjunto de espectros con buena calidad, pero únicamente una de cada 1.000 veces (0,001) en el resto de espectros de peor calidad, la probabilidad promedio asociada a esta puntuación es la combinación ponderada de probabilidades de cada uno de los espectros de alta calidad, esto es, $\mathbf{I_N(x)}$ = 0,1 x 0,2 + 0,9 x 0,001 = 0,0209. Sin embargo, si esta puntuación se obtiene a partir de un espectro perteneciente al subconjunto de alta calidad, la probabilidad promedio está subestimada por un factor de 0,1 con respecto a la probabilidad real de encontrar la puntuación al azar, que sería 0,2. Este fácil ejemplo permite darnos cuenta de los grandes errores que se pueden cometer cuando la significatividad estadística se determina utilizando la probabilidad promedio, aunque es importante destacar que los errores cometidos implican sólo una pérdida de poder discriminativo en el número total de péptidos identificados, puesto que el rigor estadístico lo sigue determinando el usuario con la tasa de error o FDR. También nos da una pista para corregir la probabilidad teniendo en cuenta el factor de calidad. Puesto que, como hemos descrito anteriormente, la probabilidad promedio de la segunda mejor puntuación es una buena estimación de la fracción de calidad, dividiendo por este número obtenemos una corrección que compensa el factor de subestimación de la probabilidad. Este razonamiento justifica cualitativamente el hecho de que la razón de probabilidad sea una buena aproximación a la probabilidad condicionada.

Esta idea se puede entender mejor considerando la Figura 17. Combinando las distribuciones de los seis espectros, construimos una distribución promedio de probabilidad. Como se muestra en A y B, dividiendo la distribución promedio de probabilidad de la mejor puntuación por la probabilidad promedio asociada a la segunda mejor puntuación obtenida en el espectro que tiene mayor calidad de los seis (el situado más a la derecha), obtenemos una distribución promedio de probabilidad corregida que, en la región de baja probabilidad, llega a ser indistinguible con la distribución de probabilidad de ese espectro en particular (línea gris). De forma similar, la probabilidad promedio correspondiente al valor más probable de la segunda mejor puntuación de cada uno de los espectros refleja muy fehacientemente la fracción de calidad de los espectros (Figura 17 C).

Propiedades de la razón de probabilidad

La razón de probabilidad se calcula en la práctica como se esquematiza en la Figura 18. La colección de espectros se busca contra una base de datos señuelo, y el conjunto de mejores puntuaciones (XCorr) se usa para construir la distribución promedio de probabilidades para la mejor puntuación $\mathbf{I_N(x)}$; esto se hace ordenando los XCorr en orden decreciente y representando

la posición normalizada en función de la puntuación. La curva resultante es la utilizada para determinar, por interpolación directa, los valores de la distribución promedio de probabilidades para la mejor y la segunda mejor puntuación, $\mathbf{I_N(x_F)}$ e $\mathbf{I_N(x_S)}$, utilizados para calcular la razón de probabilidad. La misma curva es utilizada para calcular la razón de probabilidad para las puntuaciones obtenidas al buscar contra una base de datos normal. Luego, cuando la mejor puntuación de la búsqueda contra la base de datos normal es mejor que la mayor puntuación en la distribución (esto ocurre cuando la asignación entre el espectro y el péptido es claramente positiva), la probabilidad no se calcula con una extrapolación, y simplemente se asume que es menor que la de la mejor puntuación obtenida en la búsqueda con la base de datos señuelo.

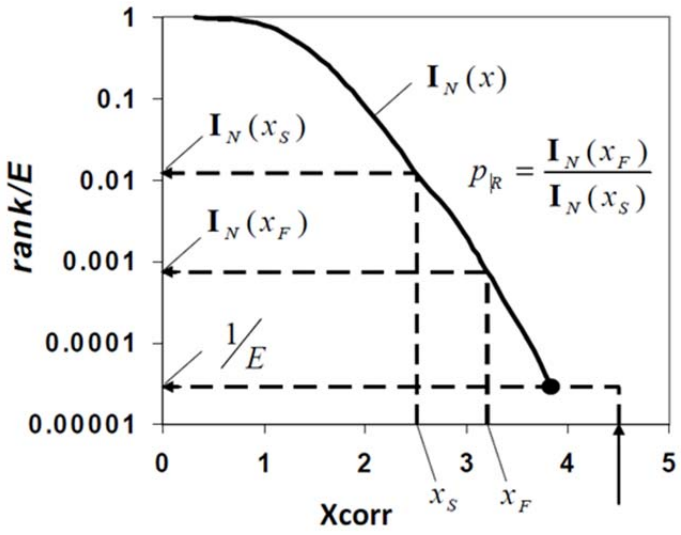


Figura 18. Cálculo de la razón de probabilidades: La distribución media de probabilidades de la mejor puntuación, $I_N(x)$, se obtiene buscando una colección suficientemente grande de espectros MS^2 contra una base de datos señuelo, y calculando para cada espectro su posición normalizada en el ranking de todos los espectros, según la mejor puntuación obtenida por cada uno de ellos. Los espectros se buscan después contra la base de datos objetivo, obteniéndose una mejor puntuación (x_F) y una segunda mejor puntuación (x_S) para cada uno de los espectros. Estas puntuaciones se interpolan numéricamente en la curva $I_N(x)$, obteniéndose las probabilidades de la mejor puntuación ($I_N(x_F)$) y de la segunda mejor puntuación ($I_N(x_S)$). La razón de probabilidades es el cociente entre estos dos valores. Cuando la mejor puntuación en la base de datos objetivo es superior a la mejor puntuación obtenida en la base de datos señuelo, únicamente se asume que la probabilidad es menor que 1/E, donde E es el número total de espectros, en este caso, 40.000; este procedimiento evita errores de estimación de probabilidades mediante extrapolación en una región cuya distribución es desconocida.

Una vez obtenidos los valores de la razón de probabilidad para la búsqueda con la base de datos normal y la señuelo, se utilizan para calcular la tasa de error o FDR ordenando las identificaciones de las dos búsquedas por el valor de la razón de probabilidad y contando la

proporción de asignaciones señuelo mejores o igual, es decir, con una razón de probabilidad menor o igual, en cada una de las asignaciones.

Como se ha explicado anteriormente la probabilidad promedio de la segunda mejor puntuación es un buen estimador de la posición relativa del espectro en la colección de acuerdo a su calidad. Por tanto, la razón de probabilidad puede concebirse como una puntuación con una corrección interna que, teniendo en cuenta la calidad del espectro, compensa la heterogeneidad de la colección de espectros MS/MS que hace que existan grandes diferencias entre un espectro y otro en términos de las puntuaciones obtenidas cuando se buscan contra una base de datos señuelo.

La razón de probabilidad puede ser por tanto calculada también como $\mathbf{I_N(x_F)/H_N(x_S)}$ construyendo de forma separada las curvas para la mejor y para la segunda mejor puntuación de la búsqueda contra la base de datos señuelo, siempre y cuando el número de secuencias candidatas es suficientemente grande. Como se muestra en la Figura 19 que muestra los resultados obtenidos para el análisis del proteoma de núcleos Jurkat, no hay diferencias apreciables entre calcular la razón de probabilidad de esta otra manera. De igual manera, la razón de probabilidad podría ser calculada usando la j-ésima mejor puntuación y la distribución promedio de probabilidades asociada a esa j-ésima puntuación.

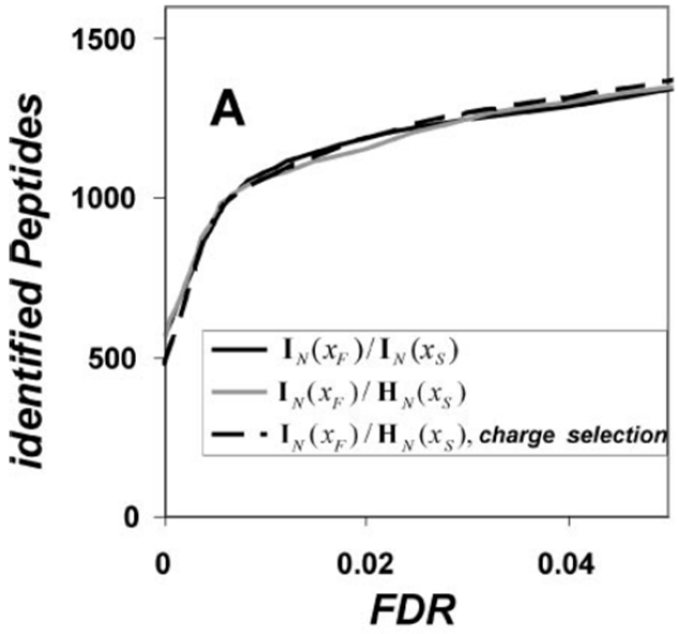


Figura 19. Efecto del método utilizado para calcular la razón de probabilidades en el rendimiento de identificación de péptidos. Una extensa colección de espectros MS^2 (más de 40.000) provenientes del análisis del proteoma de células Jurkat se sometió a un proceso de búsqueda contra bases de datos objetivo y señuelo. Los resultados de las búsquedas fueron analizados mediante el método de la razón de probabilidades, y se evaluó el rendimiento de la identificación de péptidos representando el número de péptidos identificados en función de la tasa de error FDR. La razón de probabilidades se calculó como $IN(x_F)/IN(x_S)$ (línea negra), o como $IN(x_F)/HN(x_S)$ (línea gris) o seleccionando los espectros con el estado de carga que obtengan la razón de probabilidades menor (línea discontinua).

Para comprobar si una distribución de calidades correspondiente a un conjunto de espectros determinado y usada para el cálculo de la razón de probabilidad puede afectar o no a la significancia estadística, se hizo la siguiente prueba: dado un conjunto de espectros determinado los valores de la razón de probabilidad asociados a las asignaciones de dichos espectros fueron calculados con la distribución promedio de la mejor puntuación obtenida del análisis de un conjunto diferente y mayor de espectros, y viceversa. Para ello se utilizaron los dos proteomas descritos anteriormente, los datos obtenidos del análisis de las células Jurkat humanas y los obtenidos del análisis de las células madre mesenquimales humanas. Como se puede ver en la Figura 20 no hay diferencia entre las curvas FDR calculadas de forma normal en los dos proteomas (líneas negras) y las curvas FDR calculadas tras interpolar los valores de las puntuaciones de un proteoma en las distribuciones promedio del otro (líneas grises). Como se puede ver en la Figura 21, las distribuciones promedio de las puntuaciones de los dos proteomas eran distintas, y pese a ello, las curvas FDR son indistinguibles.

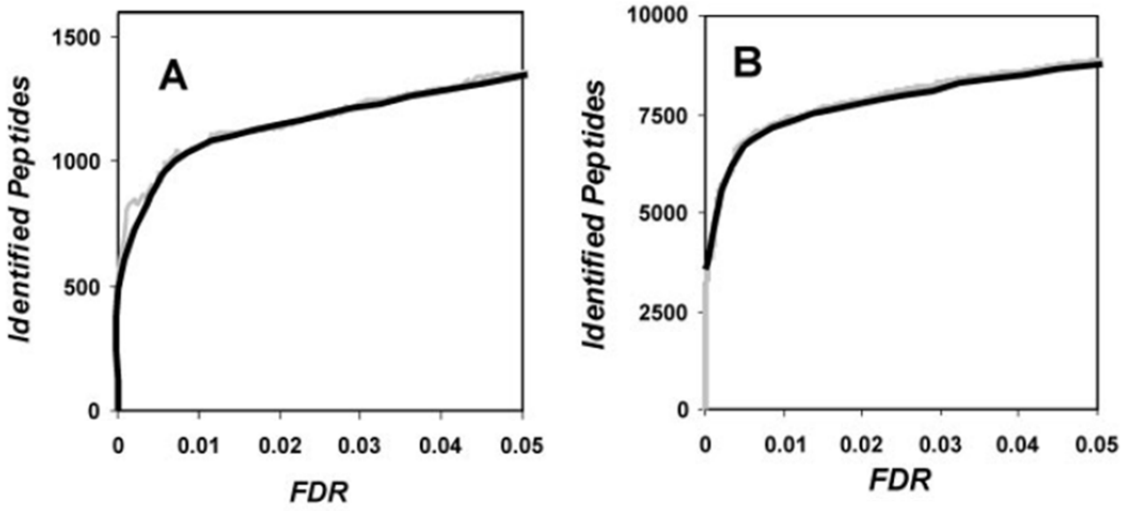


Figura 20. Efecto de las distribuciones promedio de puntuaciones usadas para calcular la razón de probabilidades en el rendimiento del método: Las líneas negras representas las curvas de FDR obtenidas del análisis de los péptidos trípticos del proteoma de células Jurkat humanas (A) y de las células madre mesenquimales humanas (B) por el método convencional del cálculo de la razón de probabilidades. Las líneas grises son las curvas de FDR obtenidas cuando las razones de probabilidades de los péptidos de un proteoma se calculan interpolando en la distribución promedio de puntuaciones del otro proteoma (A) y viceversa (B).

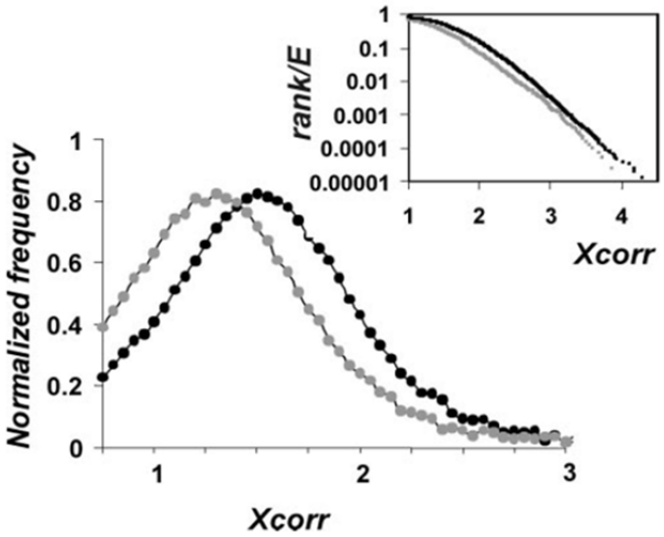


Figura 21. Comparación de las distribuciones promedio de las mejores puntuaciones de dos conjuntos de datos diferentes: Las distribuciones de las puntuaciones del proteoma nuclear (puntos grises) y del proteoma de las células madre (puntos negros) se muestran usando una frecuencia normalizada y usando en escala logarítmica (cuadro interior).

Estos resultados demuestran la robustez de la corrección de calidad introducida por la razón de probabilidad, lo que hace que los resultados finales sean independientes de la distribución de espectros MS/MS utilizada para construir la distribución promedio de puntuaciones. Nótese que

aunque se pueden utilizar diferentes distribuciones promedio para calcular los valores de la razón de probabilidad, las tasas de error deben ser siempre calculadas usando los valores de la razón de probabilidad obtenidas tras buscar el mismo conjunto de datos contra la correspondiente base de datos señuelo e interpolando en la misma distribución de puntuaciones.

Dado que la razón de probabilidad introduce una corrección intrínseca de la calidad del espectro, se quiso comprobar cómo influían otros factores que se sabe que influyen en la distribución de puntuaciones de SEQUEST. Uno de ellos es la incertidumbre acerca de la asignación del estado de carga, por el cual muchas veces un mismo espectro es buscado asumiendo diferentes estados de carga, y por tanto, es como si se considerasen espectros diferentes, lo cual puede afectar a la determinación de la FDR ya que afecta al tamaño del espacio de búsqueda. Este efecto fue probado comparando los valores obtenidos de la razón de probabilidad por el mismo espectro con diferentes estados de carga, considerándolos como si fuesen espectros diferentes y cogiendo únicamente el valor de la razón de probabilidad mejor de entre las diferentes asignaciones de estado de carga. Como se puede apreciar en la Figura 19 (línea negra discontinua), con la selección de la mejor carga la curva de FDR no varía para valores menores de 0,05 y únicamente cambia apreciablemente para valores mayores a 0,2 (no mostrado). Por tanto, una asignación correcta de la carga del péptido no es un factor crítico para la identificación del mismo cuando la razón de probabilidad se usa como indicador discriminatorio.

Sin embargo, tanto el estado de carga como la longitud del péptido son factores que afectan directamente a los valores de XCorr. Como se muestra en la Figura 22 A y B, hay un desplazamiento apreciable hacia valores más altos de XCorr cuando la longitud del péptido es mayor o cuando el estado de carga es mayor, tal y como observaron otros trabajos (Keller, Nesvizhskii et al. 2002, Nesvizhskii, Roos et al. 2006). Usando la razón de probabilidades, y clasificando de la misma manera los espectros en dichas categorías (por carga y por longitud de péptidos) las distribuciones de razones de probabilidades se centran en un mismo punto (Figura 23 A y B), corrigiéndose el desplazamiento observado con las distribuciones de puntuaciones Xcorr de la Figura 22, sin embargo se siguen observando diferencias en la forma de las distribuciones.

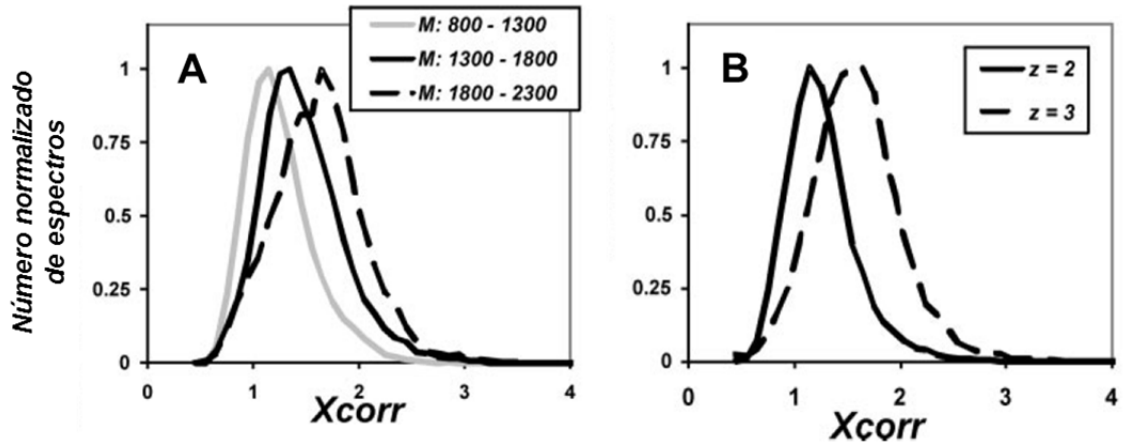


Figura 22. Efecto de la carga y la longitud del péptido en la distribución del XCorr: La colección de espectros de núcleos de células Jurkat se buscaron contra una base de datos señuelo, y los resultados se utilizaron para construir los histogramas del número total de espectros con respecto a sus respectivos valores de XCorr. En A, los espectros fueron clasificados en función de la masa del péptido (M) en tres categorías: de 8.00 a 1.300 Da (línea gris), de 1.300 a 1.800 Da (línea negra) y de 1.800 a 2.300 Da (línea discontinua). En B, los espectros fueron clasificados en función del estado de carga (z) en dos categorías: carga 2 (línea negra) y carga 3 (línea discontinua).

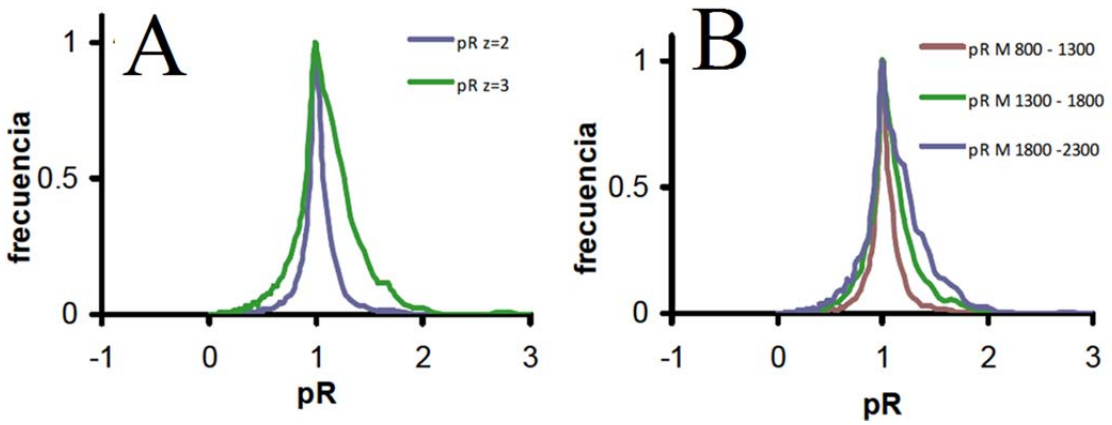


Figura 23. Efecto de la carga y la longitud del péptido en la distribución de la razón de probabilidades (PRatio): La colección de espectros de núcleos de células Jurkat se buscaron contra una base de datos señuelo, y los resultados se utilizaron para construir los histogramas del número total de espectros con respecto a sus respectivos valores de PRatio. En A, los espectros fueron clasificados en función del estado de carga (z) en dos categorías: carga 2 (línea morada) y carga 3 (línea verde). En B, los espectros fueron clasificados en función de la masa del péptido (M) en tres categorías: de 800 a 1.300 Da (línea marrón), de 1.300 a 1.800 Da (línea verde) y de 1.800 a 2.300 Da (línea morada).

Mejora de la razón de probabilidad

Un efecto observado por varios trabajos anteriores (Keller, Nesvizhskii et al. 2002, Nesvizhskii, Roos et al. 2006) es que la distribución de las puntuaciones de SEQUEST tiene una fuerte dependencia de la longitud del péptido y la carga del ion. Esto afecta a las posiciones en el ranking de las puntuaciones de SEQUEST. Este efecto altera, como puede apreciarse en la Figura 23 la cola de las distribuciones de la razón de probabilidad. Posteriormente a este trabajo, Pedro Navarro trabajó en una mejora de dicho factor como se muestra brevemente aquí.

Para ello se introdujo un factor de corrección en función de la carga (z) y de la masa (M) del ion parental. La corrección se aplicó transformando el valor de XCorr ofrecido por SEQUEST de la siguiente manera:

$$Xcorr_T(Xcorr, M, R) = \frac{log_{10}(Xcorr/R)}{log_{10}(2*M/M_0)}$$

$$donde\ \ R = \begin{cases} 1, & si\ z = 2 \\ 1.22, & si\ z = 3 \end{cases}$$

donde $XCorr_T$ es la puntuación SEQUEST corregida y M_0 es la masa promedio de los aminoácidos (M_0 = 110 Da). En esta ecuación la corrección en masa es una normalización logarítmica de la longitud del péptido estimada mediante M/M_0 y está basada en la corrección empírica sugerida por Nesvizhskii y colaboradores (Nesvizhskii, Roos et al. 2006).

En cuanto a la corrección de carga, se tuvo en cuenta el rango de masa/carga en el que la señal de ruido de fondo del espectro MS/MS pudiera ser erróneamente asignada por el motor de búsqueda a alguno de los fragmentos teóricos del péptido. Teniendo en cuenta únicamente los iones con carga 2 y con carga 3, se introdujo el factor de corrección 1,22 que corresponde a la relación entre los rangos de masas en los que iones triplemente cargados y los iones doblemente cargados pueden generar fragmentos.

Las distribuciones de las razones de probabilidades corregidas para diferentes estados de carga y longitudes de péptidos son prácticamente indistinguibles como se puede observar en la Figura 24 A y B, corrigiéndose las diferencias que se apreciaban en la Figura 23.

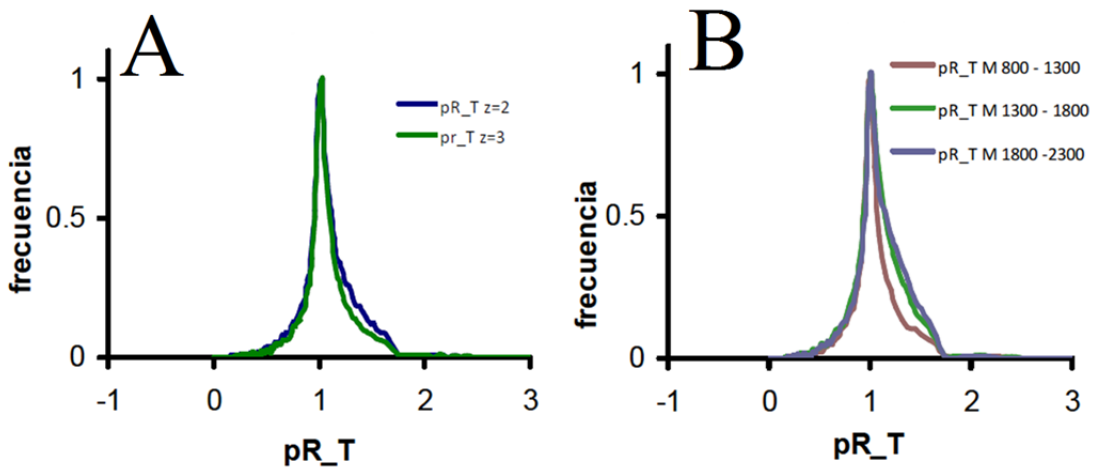


Figura 24 Efecto de la carga y la longitud del péptido en la distribución de la razón de probabilidades (PRatio) mejorada: La colección de espectros de núcleos de células Jurkat se buscaron contra una base de datos señuelo, y los resultados se utilizaron para construir los histogramas del número total de espectros con respecto a sus respectivos valores de PRatio corregidos según se describe en el texto. En A, los espectros fueron clasificados en función del estado de carga (z) en dos categorías: carga 2 (línea morada) y carga 3 (línea verde). En B, los espectros fueron clasificados en función de la masa del péptido (M) en tres categorías: de 800 a 1.300 Da (línea marrón), de 1.300 a 1.800 Da (línea verde) y de 1.800 a 2.300 Da (línea morada)

Es más, como se muestra en la Figura 25, la curva de la FDR es claramente superior cuando no se realiza ningún tipo de clasificación (diferencia entre las líneas gris y negra continuas). Como esperábamos, separando los espectros de acuerdo a sus estados de carga no resultó en una mejora significativa en comparación con el análisis realizado cuando tomamos todos los espectros juntos y la razón de probabilidad corregida es utilizada como indicador.

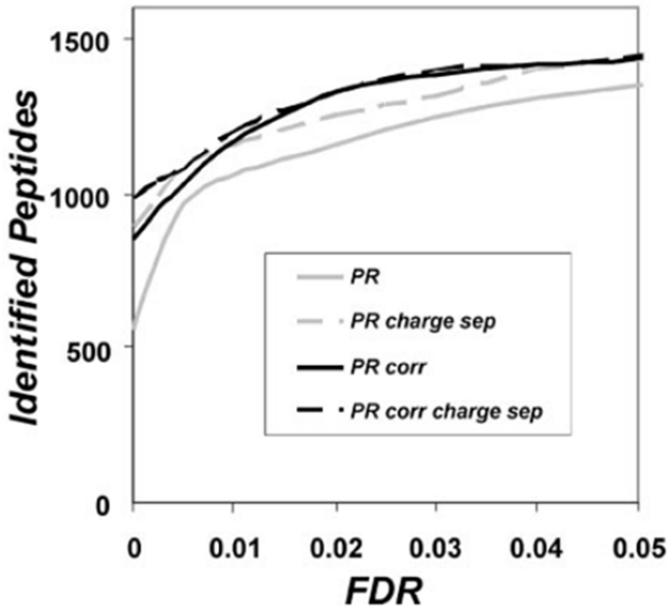


Figura 25. Efecto del método utilizado para calcular la razón de probabilidades en el rendimiento de identificación de péptidos. Una extensa colección de espectros MS^2 (más de 40.000) provenientes del análisis del proteoma de células Jurkat se sometió a un proceso de búsqueda contra bases de datos objetivo y señuelo. Los resultados de las búsquedas fueron analizados mediante el método de la razón de probabilidades, y se evaluó el rendimiento de la identificación de péptidos representando el número de péptidos identificados en función de la tasa de error FDR. Con línea gris (PR) se muestra el rendimiento calculando la razón de probabilidades como se muestra en la Figura 12. Con la línea negra discontinua se muestra la FDR cuando los espectros se clasifican de acuerdo a su estado de carga y la FDR de cada grupo de calcula separadamente (PR charge sep). Con la línea negra, el rendimiento de la FDR cuando la razón de probabilidades se corrige como se ha descrito en el texto (PR corr). Con la línea gris discontinua se muestra el rendimiento de la FDR tras corregir la razón de probabilidades y clasificando los espectros por su estado de carga (PR corr charge sep).

En la Figura 26 comparamos el método de la distribución gaussiana en dos dimensiones (2VGM) (Lopez-Ferrer, Martinez-Bartolome et al. 2004) con el uso del indicador de la razón de probabilidades con y sin la corrección. Se puede observar cómo el rendimiento del método de la distribución gaussiana y de la pRatio sin la corrección es similar. Sin embargo, comparándolo con la pRatio con la corrección el rendimiento de ésta es mucho mayor que los otros dos métodos.

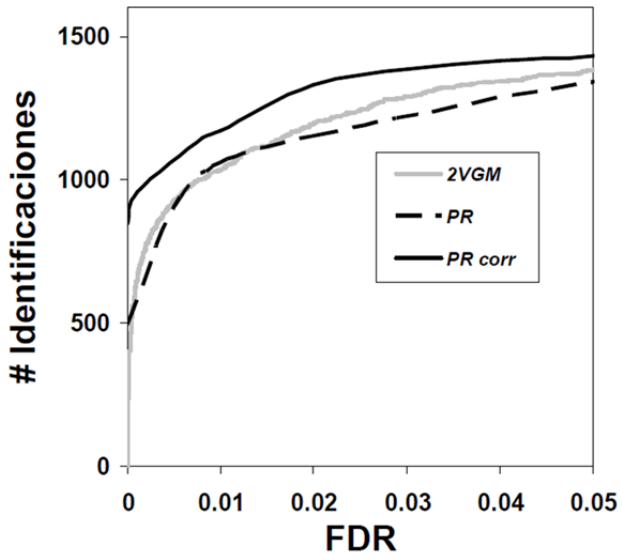


Figura 26. Comparativa del rendimiento de identificaciones obtenido usando diferentes indicadores. La misma colección de espectros MS^2 (más de 40.000) provenientes del análisis del proteoma de células Jurkat se sometió a un proceso de búsqueda contra bases de datos objetivo y señuelo. Se representó el número de péptidos identificados frente a la tasa de error de los resultados obtenidos usando como criterio cada uno de los indicadores siguientes: (gris) método de la distribución gaussiana en dos dimensiones (2VGM), (negro, línea discontinua) PR calculada usando puntuaciones XCorr no corregidas por el efecto de la masa y de la carga, y (negro, línea continua) PR usando puntuaciones XCorr corregidas.

4.1.4. El método de la calidad única

El método de la razón de probabilidad introduce, como hemos visto, una corrección relacionada con la determinación de una calidad subyacente derivada del análisis de la segunda mejor puntuación. Así pues, nos planteamos si era posible hacer estimaciones más precisas de las calidades de los espectros teniendo en cuenta más información de los resultados de la búsqueda en la base de datos.

Para ello, describimos un método simple basado en esta idea. En este procedimiento que llamamos método de calidad única, los resultados obtenidos para cada uno de los espectros se consideran por separado; la distribución de probabilidad de una puntuación, $I(x,Q)$ se estima a partir de los 1.000 primeras mejores puntuaciones obtenidas al buscar cada espectro en particular contra una base de datos, obteniéndose una distribución diferente para cada uno de los espectros. En otras palabras, se asume que cada espectro tiene una calidad diferente, así que el número de componentes de calidad en la colección es igual al número de espectros que hay en dicha colección. Las distribuciones de las mejores puntuaciones $I_N(x,Q)$ de cada uno de los espectros se determinan considerando el número N total de secuencias peptídicas candidatas contra las cuales se buscan cada uno de los espectros y aplicando la ecuación de escalado.

Luego, estas distribuciones se utilizan para calcular la significatividad estadística de la mejor puntuación de cada espectro.

Este procedimiento tiene la ventaja de que en la práctica no es necesario buscar contra bases de datos aleatorias, ya que la práctica totalidad de las mejores 1.000 puntuaciones se espera que sean al azar, incluso aunque la búsqueda se haya hecho contra una base de datos real con secuencias no modificadas.

En la Figura 14 del apartado de la ecuación de escalado podemos ver algunos ejemplos de estas distribuciones obtenidas por esta aproximación. Como se muestra, y a pesar de la simplicidad del método, las distribuciones de probabilidad estimadas para las 1.000 mejores puntuaciones, el número total de secuencias candidatas y la ecuación de escalado, se aproximan suficientemente a las distribuciones $I_N(x,Q)$ obtenidas por Monte Carlo. En la práctica, el parecido entre las distribuciones reales y las estimadas es suficientemente bueno para permitir una estimación razonable y conservativa de la FDR sin uso de bases de datos aleatorias.

En el caso de que se conozca a priori la forma analítica exacta de la distribución $I_N(x,Q)$, se pueden utilizar las puntuaciones producidas por las secuencias candidatas para calcular el valor de los parámetros que ajusten óptimamente la forma analítica a la distribución experimental. La dificultad de este método estriba en la derivación de la forma analítica de $I_N(x,Q)$, pero una vez ésta se conoce, la estimación de la confianza de las identificaciones se puede considerar un problema resuelto. En ese caso, el valor de $I_N(x,Q)$ de la mejor puntuación observada se puede extrapolar de forma fiable a partir de las puntuaciones obtenidas de todas las secuencias candidatas. Derivar la forma analítica de la distribución $I_N(x,Q)$ puede resultar especialmente difícil en muchos casos. Este cálculo depende completamente del esquema de puntuación y por tanto del motor de búsqueda y por ello no lo hemos abordado en este estudio, ya que únicamente estamos interesados en las propiedades generales aplicables a todas las distribuciones de puntuaciones. Esta aproximación se ha utilizado para calcular los valores de probabilidad con SEQUEST.

La Figura 27 muestra los resultados obtenidos con la aplicación de este método a las colecciones de espectros obtenidos del análisis de los dos proteomas, siendo en total, 40.000 (A) y 150.000 (B) espectros de fragmentación. Como se puede observar en la figura, los valores de las mejores puntuaciones estimadas, calculadas a partir de las distribuciones de probabilidad ajustadas y del número de secuencias candidatas en cada caso, se ajustan correctamente a los valores reales de las mejores puntuaciones obtenidas buscando contra una base de datos aleatoria. Este resultado nos permite saber que, en general, las distribuciones estimadas están correctamente centradas alrededor de los valores esperados.

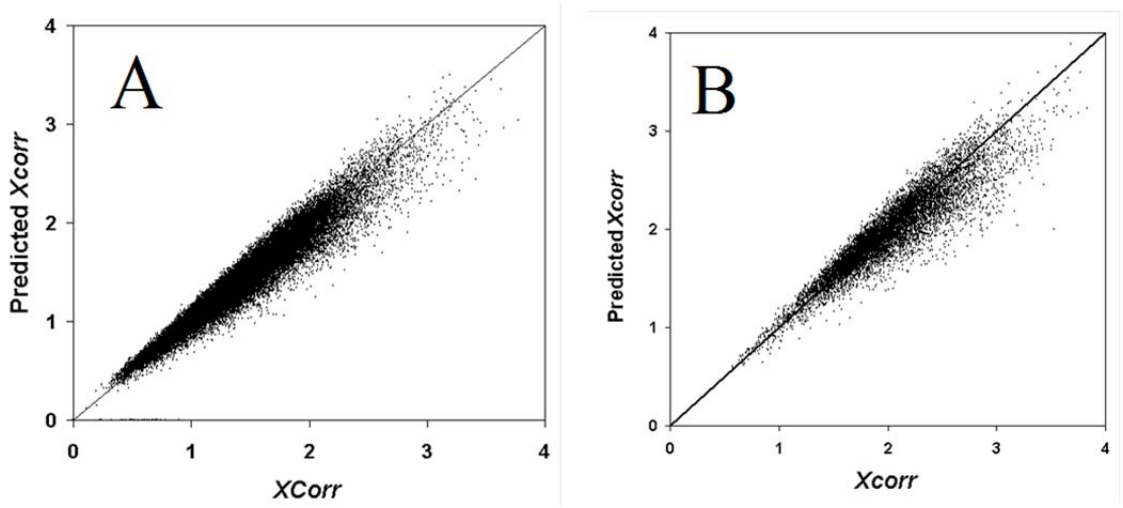


Figura 27. Pruebas del método de calidad única para estimar la distribución de la mejor puntuación de SEQUEST para un espectro de fragmentación individual. Se buscaron dos colecciones de espectros, obtenidos del análisis de dos proteomas modelo, con 40.000 (A) y 150.000 (B) espectros respectivamente, contra la base de datos humana invertida. Los valores más probables de las mejores puntuaciones estimados por las distribuciones de probabilidad extrapoladas de cada espectro contra las mejores puntuaciones observadas.

Luego, ordenando las mejores puntuaciones observadas por orden de su propia probabilidad estimada, se observa que el valor normalizado de la frecuencia de obtener una puntuación con una probabilidad igual o menor que *p* se ajusta también bastante bien a los valores de *p*, excepto por una ligera desviación negativa (Figura 28). Sin embargo, esta pequeña desviación es conservativa, es decir, la probabilidad estimada nunca es menor que la frecuencia observada, y por tanto, estas distribuciones de puntuaciones se pueden utilizar en la práctica como funciones de probabilidad, permitiendo el cálculo directo de la FDR, evitando así repetir las búsquedas contra una base de datos señuelo para calcularla.

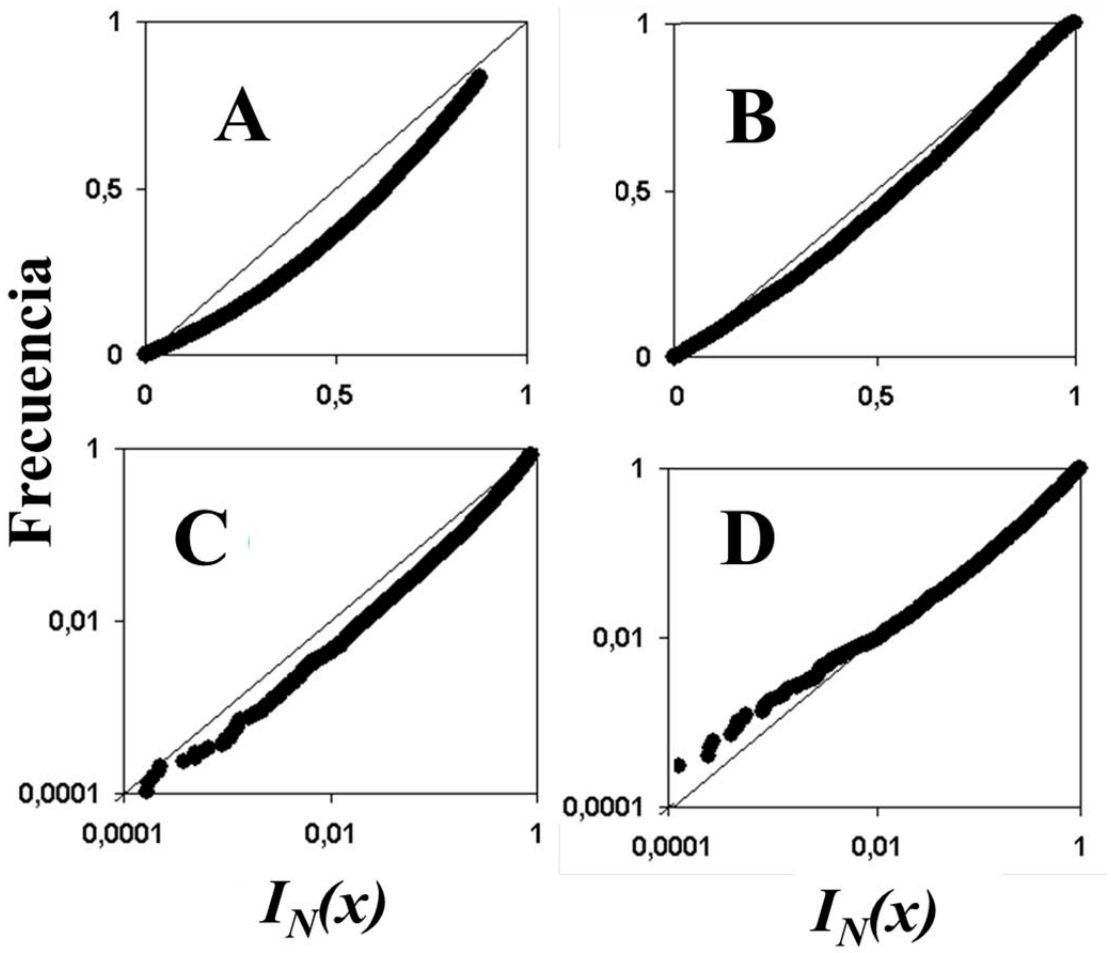


Figura 28. Pruebas del método de calidad única para estimar la distribución de la mejor puntuación de SEQUEST para un espectro de fragmentación individual. Se buscaron dos colecciones de espectros, obtenidos del análisis de dos proteomas modelo, con 40.000 (A y C) y 150.000 (B y D) espectros respectivamente, contra la base de datos humana invertida. Se representa la frecuencia acumulativa normalizada de los espectros presentados de manera normal (A y B) o logarítmica (C y D), ordenados de acuerdo a la probabilidad de sus mejores puntuaciones, en función de la probabilidad de la mejor puntuación estimada usando el método de la calidad única.

4.1.5. Prestaciones de los métodos de la razón de probabilidad y de la calidad única

Como ya se ha comentado en el texto, las correcciones aplicadas al método de la razón de probabilidad fueron realizadas por Pedro Navarro posteriormente al trabajo que aquí se presenta. El método de la razón de probabilidad, junto con las correcciones posteriores fue publicado en (Martinez-Bartolome, Navarro et al. 2008), y el método de la calidad única fue descrito de manera muy breve en dicha publicación como un método que utilizaba las distribuciones de espectro individuales.

En una comparativa inicial se muestran las prestaciones que tanto el método de la razón de probabilidad (sin la corrección), como el método de la calidad única tuvieron al analizar los datos de dos proteomas modelo (proteoma de células madre mesenquimales de médula ósea humana (A) y proteoma de núcleos de células tipo T Jurkat humanas (B)), con respecto al rendimiento obtenido analizando el mismo conjunto de datos por un método publicado anteriormente por nuestro grupo, basado en una aproximación empírica que describe el comportamiento de la mejor puntuación y de la puntuación delta (segunda mejor) en base a un modelo Gaussiano de dos variables (2VGM) (Lopez-Ferrer, Martinez-Bartolome et al. 2004), y de otro método empírico globalmente utilizado, el *Peptide-Prophet (PP)* (Keller, Nesvizhskii et al. 2002).

Como se puede ver en la Figura 29, los dos nuevos métodos (la razón de probabilidades modificada, PRM y la calidad única, SQM) desarrollados de manera analítica obtienen mejor rendimiento que los otros dos métodos descritos anteriormente (2VGM y PP). En el caso del método de la calidad única, se mostró su rendimiento de dos maneras distintas, calculando las distribuciones individuales con la base de datos normal e inversa de forma separada (SQM-DI) o únicamente con la base de datos normal (SQM-D). En este último caso, el rendimiento de dicho método se vio ligeramente reducido.

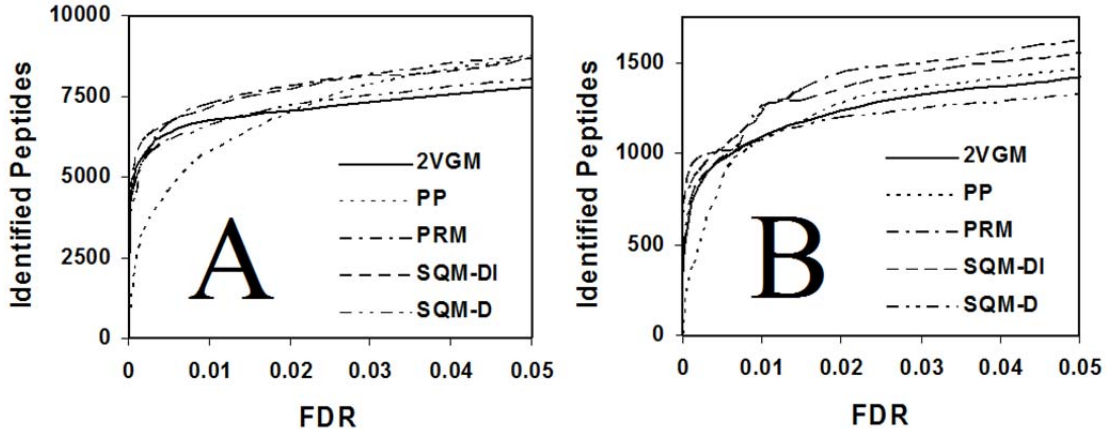


Figura 29. Comparación del funcionamiento de varios métodos estadísticos para el análisis a gran escala de los resultados de SEQUEST obtenido a partir del análisis de dos proteomas modelo. Se muestra el número de péptidos identificados en función de la tasa de error o FDR obtenidas aplicando los diferentes métodos al análisis de los resultados de buscar en una base de datos la colección total de espectros de fragmentación. Los resultados obtenidos usando los métodos empíricos de Keller et al. (PeptideProphet) (PP), y de López et al. (2VGM), se comparan con los obtenidos usando el método de la razón de probabilidad (PRM) y el método de la calidad única (SQM). En este último caso, la FDR se calculó repitiendo la búsqueda contra la base de datos invertida (SQM-DI) y directamente (SQM-D) como se explica en el texto.

Una vez desarrollada la corrección del método de la razón de probabilidad, se realizó otra comparación, esta es, la publicada en (Martinez-Bartolome, Navarro et al. 2008), analizando los mismos espectros del proteoma modelo de Jurkat, donde sí se muestra la versión corregida del método de la razón de probabilidad y en este caso los resultados obtenidos por el método de la calidad única se muestran como SSD (*single spectrum distributions*), calculadas únicamente con la base de datos normal. Como se puede ver en la Figura 30, el método de la calidad única obtiene un rendimiento similar al método de la distribución gaussiana de dos variables (2VGM). Sin embargo, incluso sin la corrección por longitud y carga, el método de la razón de probabilidad (PR) tiene un rendimiento similar al modelo de la Gaussiana de dos variables (2VGM) (el cual usa distribuciones diferentes para espectros con diferentes estados de carga), mientras que el método de la razón de probabilidad corregido tiene claramente un rendimiento superior (PR corr). Por tanto, queda demostrado que los resultados obtenidos por el método de la razón de probabilidad, desarrollado mediante consideraciones puramente analíticas y pese a su simplicidad conceptual y computacional, con ausencia de funciones o parámetros que ajustar, o clasificación de espectros, son superiores que los obtenidos por métodos empíricos diseñados específicamente para analizar los parámetros de puntuación de SEQUEST.

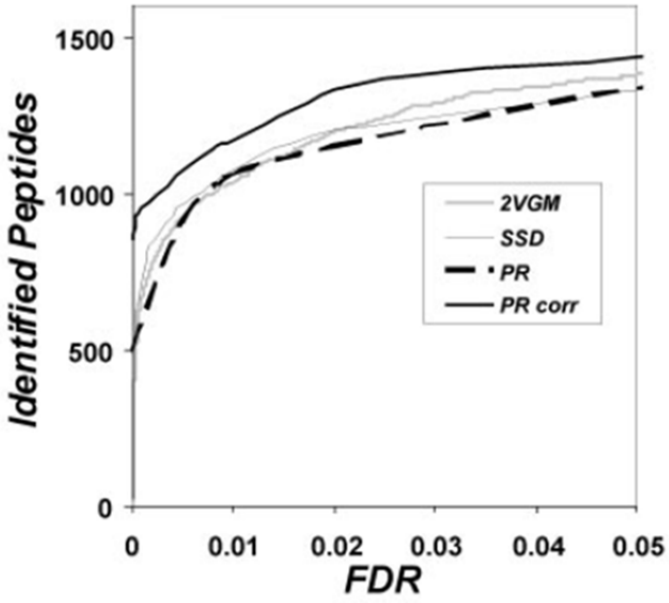


Figura 30. Comparativa del rendimiento del método de la razón de probabilidades y otros métodos empíricos: La colección de espectros MS^2 (más de 40.000) provenientes del análisis del proteoma de células Jurkat se sometió a un proceso de búsqueda contra bases de datos objetivo y señuelo. La figura presenta las curvas FDR obtenidas tras analizar las puntuaciones de SEQUEST usando el modelo Gaussiano de dos variables (2VGM; línea gris), las distribuciones de espectro único (SSD; línea gris fina), y el método de la razón de probabilidades sin corrección (PR; línea negra discontinua) y con la corrección por carga y longitud descrita (PR corr; línea negra discontinua).

Por último, también se compara el rendimiento del método de la razón de probabilidad con el obtenido usando criterios de corte fijos sobre las puntuaciones XCorr y ΔC_n, dinámicamente ajustados para obtener la tasa de error o FDR deseada (Elias y Gygi 2007). Para ello, se buscó el mismo conjunto de espectros contra una base de datos señuelo concatenada con la base de datos normal, y duplicando el número de asignaciones falsas a la hora de calcular la tasa de falsos positivos (FDR) como se sugiere en el trabajo de Elias y Gygi. Así pues, los cortes de XCorr y ΔC_n fueron iterativamente modificados hasta obtener un rendimiento óptimo en tasas de error de 0.5%, 1% y 7%, separando además los espectros en dos grupos por estado de carga, sumando luego el número de identificaciones de cada grupo en cada valor de FDR fijado. En la Figura 31 se muestran las curvas de FDR obtenidas fijando un corte óptimo de ΔC_n para cada estado de carga y valor de FDR, y variando luego el corte por XCorr. Como se puede observar, pese a que las curvas obtenidas por esos métodos se aproximan a la curva de la razón de probabilidad corregida, el rendimiento general es en todos los casos peor. Además la clasificación en grupos por estado de carga es un paso crítico para esta metodología, como se puede observar en el rendimiento claramente inferior que se obtiene cuando no hacemos esta clasificación (línea discontinua fina), lo cual en el método de la razón de probabilidades se realiza de una forma interna con la corrección por carga y masa del XCorr, sin necesidad de utilizar distribuciones separadas.

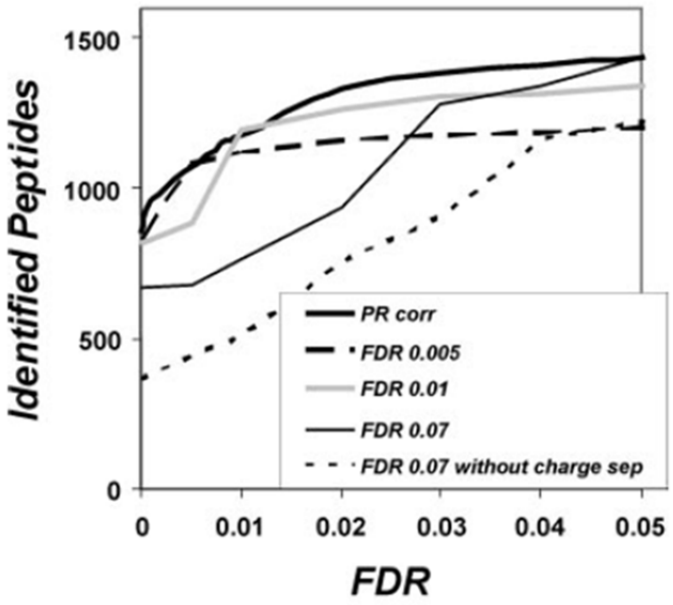


Figura 31. Comparativa del rendimiento entre el método de la razón de probabilidades y el método de optimización de cortes de XCorr y ΔCn: La curva de FDR obtenida con el método corregido de la razón de probabilidades (PR corr; línea negra gruesa) se comparó con las curvas obtenidas al aplicar criterios de corte sobre XCorr y deltaCn, los cuales fueron iterativamente modificados para obtener el máximo número de péptidos identificados a valores de FDR fijos, usando una base de datos señuelo concatenada con una base de datos normal. Los espectros fueron separados en dos grupos por estado de carga y los criterios de corte fueron optimizados para las FDRs de 0,005 (FDR 0,005; línea discontinua), 0,01 (FDR 0,01; línea gris) y 0,07 (FDR 0,07; línea negra fina), además de a 0,07 pero sin separar los espectros por estado de carga (FDR 0,07 without charge sep; línea discontinua fina).

En el trabajo de Elias y Gygi (Elias y Gygi 2007) se sugiere que la estimación de la tasa de error con el uso de bases de datos señuelo es mejor utilizando una base de datos normal concatenada con la base de datos señuelo (construida a partir de la normal) que realizando de forma separada en dos búsquedas independientes. Buscar con la base de datos concatenada permite la competición directa entre las secuencias de la base de datos normal y la señuelo por las mejores puntuaciones, con lo que se evita el hecho frecuente de obtener puntuaciones relativamente altas en asignaciones de espectros MS/MS que realmente corresponden a péptidos reales con secuencias peptídicas de la base de datos señuelo. En otras palabras, el uso de bases de datos señuelo concatenadas con sus correspondientes bases de datos normales, evita el efecto de calidad producido por los mejores espectros MS/MS. En el caso de la razón de probabilidad, aunque las búsquedas están hechas en bases de datos separadas, el efecto de la calidad de los espectros se tiene en cuenta intrínsecamente y por tanto dichos efectos son inapreciables. Para demostrar esto, los valores de FDR obtenidos mediante el método de optimización de puntuaciones de corte de SEQUEST usando una base de datos concatenada se compararon con los valores de FDR obtenidos usando dos búsquedas independientes, una sobre la base de datos

normal y otra sobre la base de datos señuelo. Lo mismo se hizo para el método de la razón de probabilidad. Como se puede ver en la Figura 32, cuando se usa el método de optimización de cortes de puntuaciones mediante búsquedas separadas (señuelo + normal), los valores de FDR se incrementan en un 40 %. Sin embargo, no se aprecian diferencias significativas cuando se hace esta comparación en el método de la razón de probabilidad.

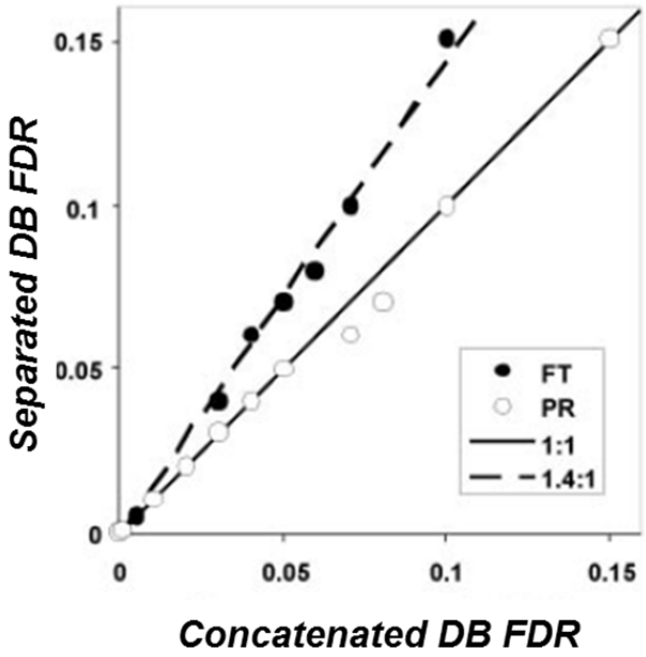


Figura 32. Comparación de las FDR obtenidas al utilizar una base de datos decoy concatenada con la base de datos normal, y al utilizar las bases de datos separadas: Se compararon los valores de FDR obtenidos usando búsquedas en bases de datos separadas con los valores de FDR obtenidos usando una base de datos concatenada, para el método de optimización de cortes de las puntuaciones de SEQUEST (FT; puntos negros) y para el método de la razón de probabilidades (PR; puntos blancos). Los números indican la pendiente de las rectas.

De la misma manera se llegó a la misma conclusión analizando otros factores que pudiesen cambiar el tamaño del espacio de búsqueda, como ciertos parámetros de la búsqueda: las modificaciones variables, la tolerancia de la masa del precursor o el máximo número de cortes enzimáticos omitidos (*missed cleavages*). Incluso, el rendimiento ofrecido por nuestro indicador fue probado también en situaciones donde el espacio de búsqueda es muy pequeño, es decir, las posibles secuencias candidatas para cada espectro no son muy numerosas. Para ello, aplicamos la razón de probabilidad corregida a una colección de 10.000 espectros MS/MS obtenidos del análisis de un extracto proteico de levadura *E. Coli* usando un espectrómetro LTQ-Orbitrap. Los datos se buscaron contra una base de datos que contenía únicamente proteínas de *E. Coli* y se utilizó una tolerancia del precursor de 10 ppm, sin modificaciones variables y sin permitir la omisión de cortes de la tripsina. El rendimiento en el número de identificaciones comparando

las curvas de FDR fue también superior con nuestro método comparado con el resto de métodos empíricos (datos no mostrados).

4.1.6. Una puntuación de probabilidad normalizada

A pesar del desarrollo técnico existente en la Proteómica moderna, no existen aún herramientas universales para validar los resultados publicados (Carr, Aebersold et al. 2004). El presente estudio proporciona varias posibilidades que podrían ayudar a definir un criterio general para validar las asignaciones de péptidos. Puesto que la ecuación de escalado se cumple en la región de posible identificación (Figura 16), se podría utilizar para calcular la probabilidad promedio normalizada de los resultados obtenidos por diferentes laboratorios o en diferentes condiciones, ya que el método equivale a igualar el espacio de búsqueda con un número predefinido de secuencias peptídicas únicas.

Este método, sin embargo, únicamente se basa en la distribución de la mejor puntuación, y, como se explicó anteriormente, resulta más que recomendable utilizar también la información que proporciona la segunda mejor puntuación. A este respecto, podemos utilizar la propiedad de que la razón de probabilidad es proporcional a N en la región de posible identificación, lo cual permite estimar la razón de probabilidad normalizada cuando se busca contra un número predefinido de secuencias candidatas. Como se muestra en la Figura 33, la probabilidad promedio de las mejores puntuaciones de un conjunto de espectros, obtenida utilizando la distribución promedio de las mejores puntuaciones al buscar contra cuatro bases de datos diferentes, es dependiente de la base de datos. Sin embargo, si utilizamos el indicador de la razón de probabilidad, el valor obtenido para todos los espectros es muy parecido independientemente de la base de datos utilizada.

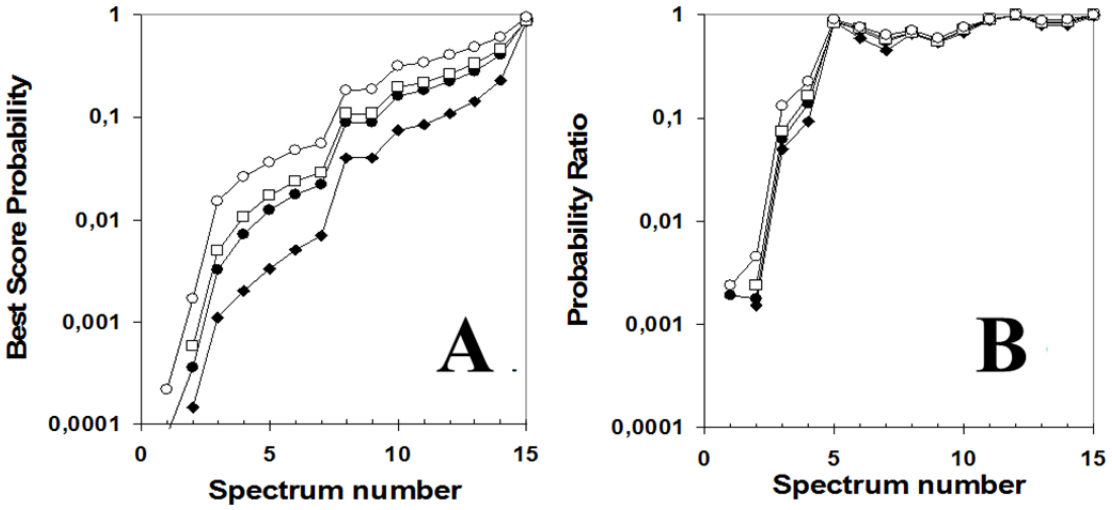


Figura 33. Hacia una probabilidad de la puntuación normalizada e independiente del experimento. Se seleccionaron aleatoriamente una colección de espectros, que luego se buscaron contra las bases de datos señuelo de diferentes tamaños: levadura (rombos negros), humana (círculos negros), swissprot (cuadrados vacíos) y nr (círculos vacíos), asumiendo una especificidad tríptica. La figura muestra la probabilidad promedio de las mejores puntuaciones (A) y las razones de probabilidad (B) obtenidas en cada caso. Los espectros se ordenaron de acuerdo a su calidad, y se unieron los puntos para una mejor claridad de la figura.

El comportamiento de las distribuciones de probabilidad obtenidas por el método de calidad única se puede predecir de manera precisa con la ecuación de escalado. Por tanto, este método debería también permitir calcular una puntuación normalizada cuando busquemos contra una base de datos con un número de secuencias candidatas predefinido. Esta idea se muestra en la Figura 34, donde se puede ver que las distribuciones de probabilidad dependen del tamaño de la base de datos, y por tanto, la mejor puntuación estimada con el método de la calidad única es diferente para un mismo espectro buscado en diferentes bases de datos (A). Sin embargo, si aplicamos la ecuación de escalado, asumiendo un *N* común, vemos que la puntuación esperada para cualquier base de datos es similar (B). Por tanto, el método de la calidad única junto con la ecuación de escalado permite una normalización efectiva de las probabilidades obtenidas con diferentes condiciones.

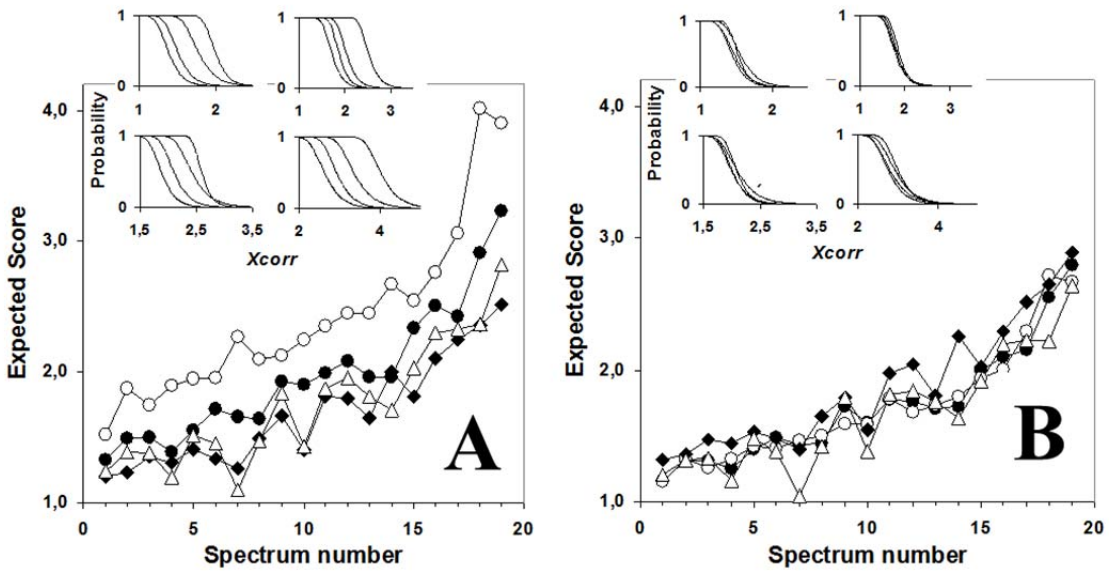


Figura 34. Hacia una probabilidad de la puntuación normalizada e independiente del experimento. Se seleccionaron aleatoriamente una colección de espectros, que luego se buscaron contra las bases de datos señuelo de diferentes tamaños: levadura (rombos negros), humana (círculos negros) y nr (círculos vacíos), asumiendo una especificidad tríptica, o buscando contra la base de datos humana asumiendo la especificidad de la EndoLysC (triángulos vacíos). En (A) se muestra el valor más probable de la mejor puntuación, determinada a partir de las distribuciones de probabilidad obtenidas para cada una de las bases de datos utilizadas. En (B) se muestran lo mismo que en (A), exceptuando que todas las distribuciones de probabilidad han sido ajustadas a 5.000 secuencias candidatas, usando la ecuación de escalado con N=5.000. Los espectros se ordenaron de acuerdo a su calidad, y se unieron los puntos para una mejor claridad de la figura.

4.1.7. Implementación de los métodos de la razón de probabilidad y de la calidad única en herramientas bioinformáticas

Como hemos visto, estos métodos están basados en la corrección de la probabilidad teniendo en cuenta la calidad subyacente a cada espectro. Dicha calidad se estima con la información que proporciona el análisis del segundo mejor score en el caso del método de la razón de probabilidad, y del análisis de los 1.000 mejores scores en el caso del método de la calidad única. Debido a su gran utilidad, desarrollamos una aplicación para cada método con el objetivo de disponer de una herramienta de libre acceso que permita una rápida y fácil aplicación de estos métodos.

La interfaz de la aplicación del método de la razón de probabilidad se muestra en la Figura 35. En la aplicación es necesario seleccionar el directorio en el que se han creado los ficheros '.out' al realizar la búsqueda de los espectros con SEQUEST contra una base de datos real y contra su correspondiente base de datos invertida.

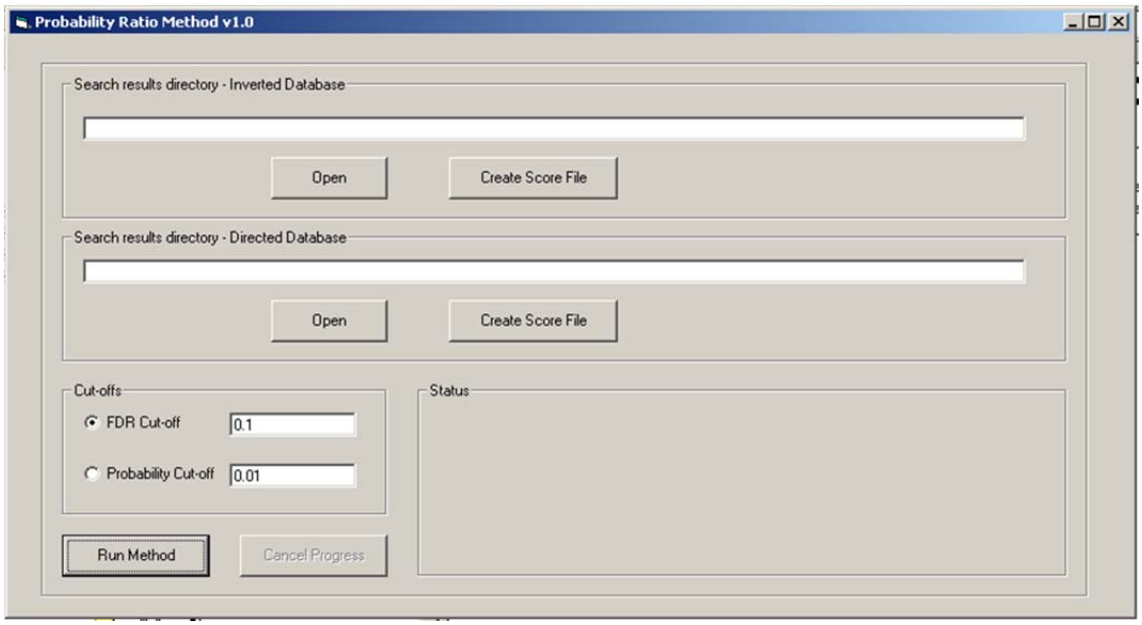


Figura 35. Interfaz gráfica de la aplicación para el método de la razón de probabilidad. El usuario tiene dos cajas de texto donde debe seleccionar los directorios donde se encuentran los ficheros ".out" de salida de las búsquedas normal y señuelo. Luego, selecciona un corte por FDR o por probabilidad, y pinchar en el botón "Run Method" para obtener la salida del programa en un fichero Excel. También puede crear un fichero "score file" para cada búsqueda, donde se recopila la información de los múltiples ficheros ".out", lo cual mejorará el rendimiento de la aplicación en futuros análisis de los mismos datos.

La salida del programa es un documento en formato de Microsoft Excel con una tabla con varias columnas que contienen información acerca de cada espectro. Esta tabla se presenta ordenada por la probabilidad de los espectros, y sólo mostrará la información de los espectros que pasen el umbral de corte definido en la interfaz del programa, teniendo la posibilidad de cortar, o bien por un umbral de probabilidad, o bien de FDR. En otra hoja del mismo documento se presenta una simple gráfica que representa la FDR frente al número de péptidos identificados.

Por otra parte, la interfaz para el método de la calidad única es bastante similar a la del método anterior, y se muestra también en la Figura 36:

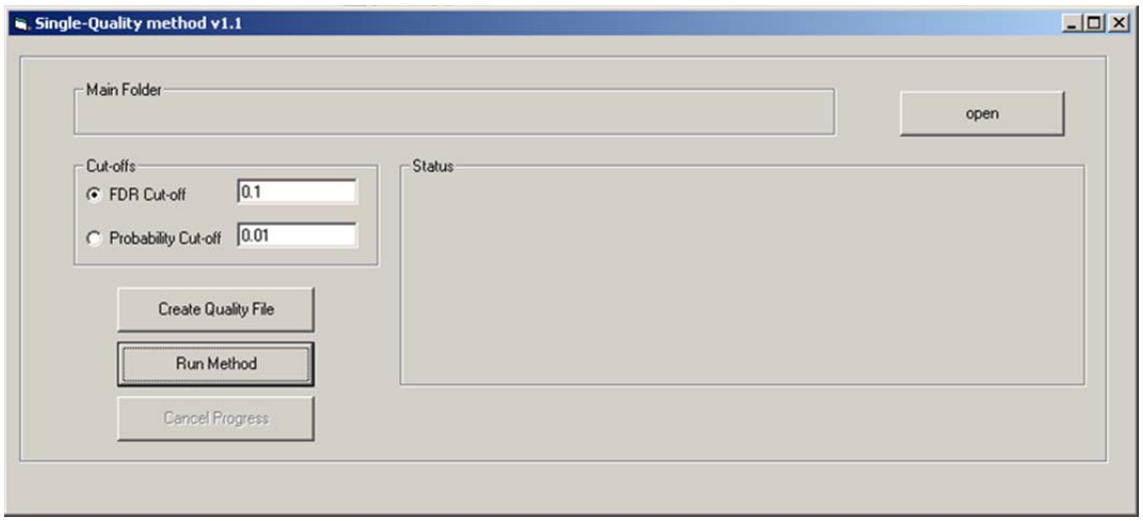


Figura 36. Interfaz gráfica de la aplicación para el método de la calidad única. En este caso el usuario sólo tiene que seleccionar el directorio donde se encuentran los ficheros ".out" de salida de la búsqueda contra la base de datos normal. Luego, selecciona un corte por FDR o por probabilidad, y pinchar en el botón "Run Method" para obtener la salida del programa en un fichero Excel. También puede crear un fichero "quality file" donde se recopila la información de los múltiples ficheros ".out", lo cual mejorará el rendimiento de la aplicación en futuros análisis de los mismos datos.

En este caso, únicamente es necesario seleccionar un directorio en el que se hayan creado los ficheros de salida de SEQUEST, ya que este método no necesita que se repita la búsqueda contra una base de datos señuelo, al conseguir la información acerca de la distribución de las puntuaciones aleatorios a partir cada espectro. Para ello, es necesario que estos ficheros de salida de SEQUEST contengan la información de las 1.000 primeras mejores puntuaciones (ver sección de materiales y métodos 3.1.6).

4.2. Desarrollo de herramientas basadas en estándares HUPO-PSI

4.2.1. Desarrollo de un repositorio online de experimentos proteómicos basados en las directrices MIAPE

Las directrices MIAPE desarrolladas por el HUPO-PSI se definen en diferentes módulos dependiendo de la técnica o del paso intermedio en el flujo de análisis proteómico que se esté describiendo. La descripción completa de un experimento deberá seguir por tanto las directrices de varios módulos MIAPE, como se detalló en el apartado 1.5.3.

Muchos investigadores proteómicos consideran una tarea demasiado costosa recopilar y escribir la información requerida por estas directrices. Lo cierto es que se requiere un tiempo considerable para entender qué información se está pidiendo y para recopilarla, sobre todo para los investigadores que no trabajan directamente con los espectrómetros de masas o que no analizan por sí mismos los resultados de éstos. Además, las herramientas de manejo de datos proteómicos como los LIMS (*Laboratory Information Management System*) disponibles no permiten extraer dicha información de forma automática.

Para resolver estos problemas desarrollamos una herramienta web, la herramienta informática generadora de documentos MIAPE, (*MIAPE Generator tool)* (Martinez-Bartolome, Medina-Aunon et al. 2010), que ayuda a sus usuarios a crear documentos MIAPE para describir sus experimentos, de una manera fácil y rápida. La herramienta está disponible para el uso de la comunidad científica desde la web de ProteoRed ISCIII (Instituto Nacional de Proteómica – Instituto de Salud Carlos III) (Paradela, Escuredo et al. 2006) o directamente en la dirección: http://www.proteored.org/MIAPEGenerator. Básicamente, proporciona una serie de formularios web y de plantillas de ejemplo para guiar a los usuarios a lo largo del proceso de creación de los documentos (Figura 37). La información requerida para cada una de las secciones MIAPE se muestra al usuario para informarle de qué tipo de dato se requiere en cada momento, siguiendo la estructura de los documentos como se muestra esquemáticamente en la Tabla 3. Una vez que el usuario ha incluido toda la información, se puede generar un informe o documento final, el cual puede adjuntarse, por ejemplo, junto con la información suplementaria asociada al manuscrito del experimento proteómico descrito. Así pues, esta herramienta permite de una manera muy simple proporcionar la información MIAPE a las revistas especializadas, con la ventaja de que de este modo se demuestra que el manuscrito tiene la información mínima exigida por las directrices MIAPE y puede ser fácilmente verificado por la revista.

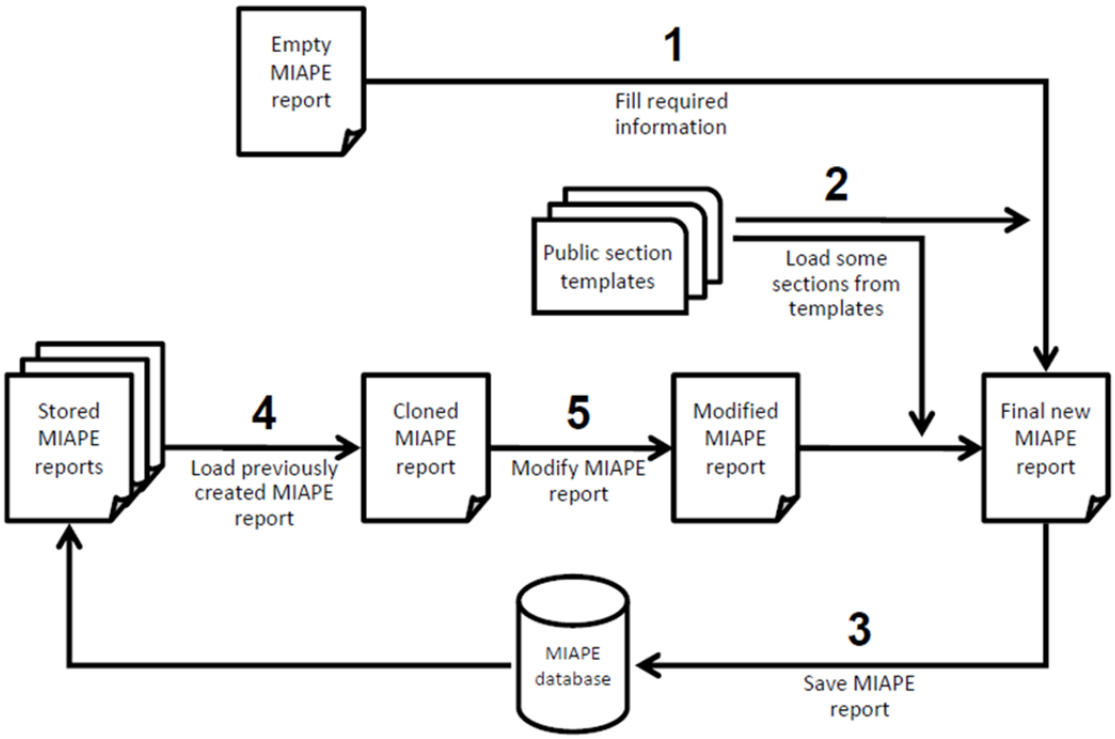


Figura 37. Flujo de trabajo de la herramienta generadora de documentos MIAPEs: (1) El usuario puede generar un documento vacío y completar la información MIAPE rellenando los formularios manualmente. (2) El usuario puede cargar también plantillas predefinidas de secciones concretas del documento, o del documento completo, para así ahorrar tiempo para completar el documento. (3) Todos los documentos creados se almacenan en la base de datos. (4) El usuario puede cargar un documento previamente creado de la base de datos, y (5) modificarlo para generar un nuevo documento.

Al ser una herramienta online, el primer paso en el flujo de trabajo es el inicio de sesión por parte del usuario, simplemente proporcionando un nombre de usuario (su correo electrónico) y una contraseña. Este usuario tendrá su propio ámbito y, por defecto, todos los documentos almacenados con su cuenta de usuario serán privados y no podrán ser consultados ni modificados por nadie más. Sin embargo, el usuario tiene la posibilidad de decidir qué información quiere compartir con otro usuario concreto conocido o hacerla totalmente pública, siempre asignando ciertos tipos de privilegios: 1) permiso de sólo lectura – otros usuarios podrán ver el documento final, pero no podrán editarlo; 2) permiso de modificación – otros usuarios pueden ver, editar, añadir y borrar información en el documento; 3) permiso para compartir – otros usuarios pueden a su vez compartir los documentos con otros. También existe un usuario "*invitado*" (*guest*), con nombre de usuario 'guest' y contraseña 'guest' que tiene visibles todos los documentos marcados como públicos por otros usuarios, y a su vez cualquier dato introducido por ese usuario será visible automáticamente por el resto. El sistema basado en ámbitos de usuarios preserva la seguridad y privacidad sobre los datos, al igual que permite la

creación de un repositorio público o privado de experimentos proteómicos perfectamente documentados.

La interfaz de usuario, como hemos dicho, está basada en una serie de formularios web para introducir los datos. Estos formularios siguen una estructura jerárquica (Figura 38 A y B), reflejando las secciones de los módulos MIAPE definidos en el HUPO-PSI (Tabla 3). Cada documento se crea siempre dentro de un proyecto y cada proyecto puede contener un número ilimitado de documentos MIAPE.

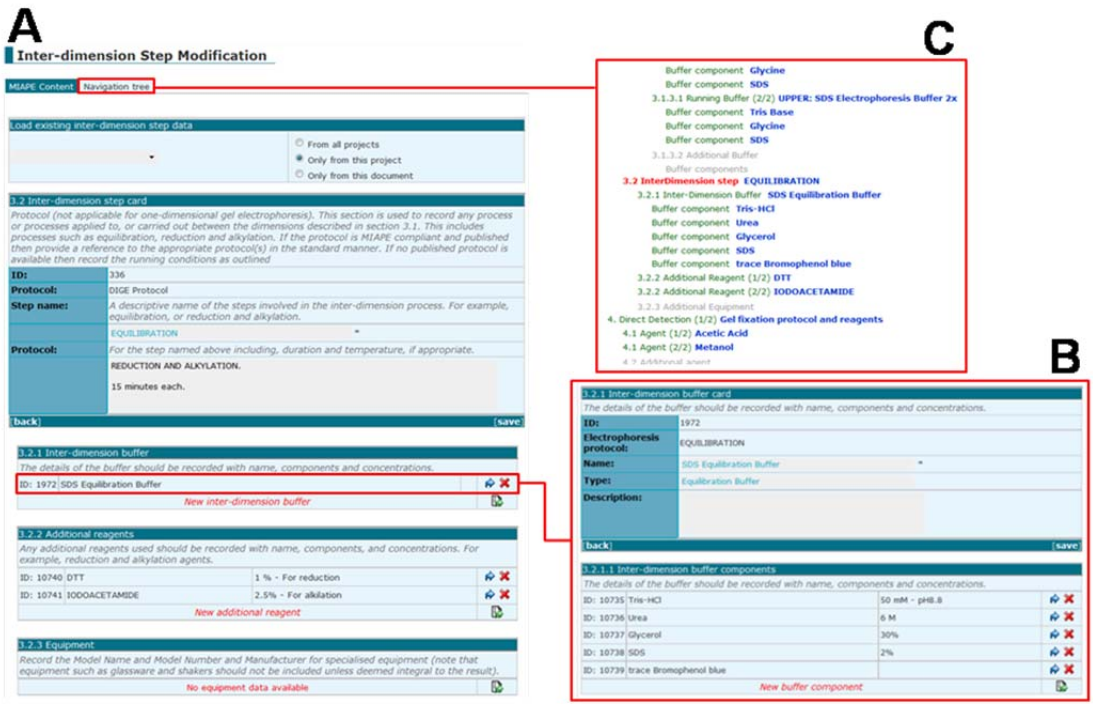


Figura 38. Interfaz web de la herramienta generadora de documentos MIAPEs: La interfaz de usuario correspondiente a la sección *'interdimension step'* de las directrices MIAPE para experimentos de electroforesis por gel. En (A) se muestran los formularios web que piden la información correspondiente a una sección del módulo MIAPE. (B) Siguiendo la estructura jerárquica del documento original, se puede acceder a sub-secciones en la parte de abajo de las páginas, pinchando en las flechas azules. (C) En todas las secciones está disponible un árbol de navegación que permite al usuario moverse entre las diferentes secciones del documento de manera directa, simplemente pinchando en su nombre. Además, esta vista de árbol proporciona información acerca de qué sección es la actual, qué secciones están ya creadas y cuáles están aún pendientes por crear.

Actualmente, la herramienta generadora de MIAPEs es capaz de crear y manejar la información de cuatro tipos de módulos MIAPE:

- MIAPE GE: *Gel Electrophoresis* (desde la elaboración del gel a la adquisición de imagen) (Gibson, Anderson et al. 2008).
- MIAPE GI: *Gel Informatics* (análisis de las imagines de los geles y el tratamiento estadístico de los datos) (Hoogland, O'Gorman et al. 2010).

- MIAPE MS: *Mass Spectrometry* (desde la adquisición de espectros a la generación de los ficheros de listas de picos) (Taylor, Binz et al. 2008).
- MIAPE MSI: *MIAPE Spectrometry Informatics* (análisis de esas listas de picos para producir identificaciones de péptidos y proteínas) (Binz, Barkovich et al. 2008).

En muchos casos, los flujos de trabajo de los laboratorios de Proteómica siguen unos protocolos predefinidos que implican el uso de ciertos instrumentos y aplicaciones bioinformáticas. La descripción de estos procesos, equipamiento, parámetros, software, etc… que se utilizan de manera rutinaria en un laboratorio y de la misma manera en diferentes experimentos, pueden ser descritos y almacenados por primera vez en la herramienta generadora de documentos MIAPE para así crear una plantilla conteniendo únicamente la información común entre experimentos. Luego esta plantilla podrá ser de gran utilidad para crear documentos MIAPE describiendo otros experimentos, teniendo sólo que introducir la información exclusiva de cada uno. Esto es posible gracias a la base de datos relacional que almacena todos los datos introducidos en una estructura jerárquica, lo que permite cargar los datos previamente creados de una sección MIAPE y todas las subsecciones que dependan de ésta para generar un documento nuevo. Esta característica, usando por ejemplo plantillas Excel, no sería posible. Gracias a la colaboración de los diferentes laboratorios de ProteoRed, se crearon una serie de plantillas tipo, accesibles a cualquier usuario, de gran utilidad para aquellos que accedan por primera vez a la herramienta o no conocedores de los detalles requeridos por las directrices MIAPE.

Un árbol de navegación, accesible en todo momento, permite a los usuarios moverse de una sección a otra del documento, evitando la necesidad de completarlo de una forma lineal. Además, la herramienta lleva a cabo una validación para informar al usuario qué secciones se han completado ya y cuáles no (Figura 38 C).

En varios estudios se ha advertido que pequeños cambios en los métodos experimentales (Turck, Falick et al. 2007) o en el procesamiento bioinformático pueden afectar tremendamente el resultado de los experimentos proteómicos (Bell, Deutsch et al. 2009). Dichas diferencias deberían ser detectadas revisando los informes conformes a las directrices MIAPE que describen la información crítica que afectará a los resultados de los experimentos. Esta revisión puede realizarse también en la herramienta aquí descrita, ya que es capaz de chequear y revisar simultáneamente información concerniente a múltiples experimentos, ofreciendo la posibilidad de compararlos mostrando en una sola tabla los protocolos e información de interés de cada uno (Figura 39). Esta funcionalidad puede ayudar a identificar los factores causantes de resultados incoherentes, como un paso particular en el protocolo de análisis, la tecnología utilizada, un parámetro del motor de búsqueda, etc…

Comparison of: Input parameters				
Export	MIAPE ID: 160	MIAPE ID: 164	MIAPE ID: 166	MIAPE ID: 168
	MIAPE MSI UVEG	LP-CSIC/UAB DIGE Proteored 2008 MSI (MALDI-TOF)	QStar_iTRAQ_UPF	UB_MS_Dige_Proteored_2008
Input parameters				
ID	188	192	195	202
Name	Search Parameters - Trypsin	Parameter Protein Prospector MS-Fit	IPI_HUMAN v3.49	Input parameters
Taxonomical restrictions	Bacteria	Mammals	74013 entries	Homo sapiens
Specified cleavage agent	Trypsin (Preferentially cleaves at Arg and Lys in position C-terminal with higher rates for Arg)	Trypsin	Trypsin	
Misscleavages	1	2	Not specified	1 missed cleavage
Permissible AA modifications	Fixed modification: Carbamidomethyl (C) Variable modifications: Oxidation (M)	Peptide N-terminal Gln to pyroGlu | Oxidation of M | Protein N-terminus Acetylated |	Not specified. The software considers more than 100 modifications	Carbamidomethyl
Thresholds; minimum scores for peptides, proteins	Maximum Peptide Rank:10 Max. number of peptides: 100 Max. number of MSMS peaks: 150		p value <0.05	p-value
Mass tolerance for PMF	Peptide tolerance: 50-100 ppm	50 - 100 ppm		50ppm
Additional parameters related to cleavage			Consideration of semi-specific cleavages	
Precursor-ion and fragment-ion mass tolerance for tandem MS			Not specified	

Figura 39. Comparación de documentos MIAPE: La tabla contiene información acerca de los parámetros de entrada utilizados en diferentes búsquedas MS/MS en bases de datos. Cada fila contiene información acerca de ciertos parámetros y cada columna contiene información procedente de cuatro documentos MIAPE diferentes, en este caso, con identificadores 188, 192, 195 y 202. La tabla puede exportarse a una hoja Excel.

La herramienta generadora de documentos MIAPE en ProteoRed

La incorporación de los estándares de HUPO-PSI como una manera común para intercambiar los datos entre los miembros de la red es uno de los objetivos de las iniciativas de estandarización de ProteoRed (Paradela, Escuredo et al. 2006, Martinez-Bartolome, Blanco et al. 2010). Así pues, algunos de sus miembros has incorporado las directrices MIAPE dentro de sus flujos de trabajo diarios, incluyendo los documentos MIAPE en los resultados enviados a los usuarios u ofertándolos como una parte adicional de los servicios.

ProteoRed ha organizado diversos experimentos multi-centro en los que todos sus miembros e incluso laboratorios externos participaban en el análisis de una misma muestra, para así tratar de resolver o de optimizar los problemas proteómicos más comunes como la

reproducibilidad en experimentos basados en geles, o de expresión diferencial, pudiendo aprender los unos de los otros de manera colaborativa. Todos estos experimentos han sido descritos siguiendo las directrices MIAPE, usando esta herramienta (hasta la aparición del MIAPE Extractor, descrito más adelante en 4.2.5). Los resultados obtenidos y los protocolos aplicados son discutidos en una puesta en común formada por todos los grupos participantes. En este contexto, la herramienta resultó ser de gran utilidad para realizar comparación entre los informes de diferentes experimentos y así poder llegar a conclusiones sobre cuál es la mejor aproximación.

Con la intención de obtener comentarios acerca de la usabilidad de la herramienta y sobre las directrices MIAPE en sí, a finales del 2008 se elaboró una encuesta dirigida a todos los laboratorios que habían usado esta herramienta. Los resultados de dicha encuesta se publicaron en la web: http://www.proteored.org/MIAPE_Survey_results_nov08.html. En ella se puede ver que la mayoría de comentarios son positivos (sólo un 5% de ellos indicaron que el sistema no era útil), y sin embargo, la mayoría (70%) indicaron que la herramienta podría proporcionar un grado adicional de calidad a sus usuarios. La encuesta también sirvió para recoger sugerencias sobre posibles mejoras. En todo caso, tanto los comentarios negativos como las sugerencias de mejora se utilizaron para desarrollar nuevas funcionalidades, y fue de gran utilidad para la iniciativa HUPO-PSI con quienes compartimos el resultado de esta encuesta, lo que suscitó el debate sobre las directrices MIAPE y ayudó a la adaptación de algunos módulos MIAPE, revisando su practicidad, y posteriormente simplificándose en parte.

Un problema detectado tras el uso de las primeras versiones de la herramienta fue la gran heterogeneidad en los textos introducidos por los usuarios, lo que dificultaba las comparativas entre experimentos cuyos informes MIAPE habían sido realizados por diferentes personas. En segundo lugar, cualquier trabajo de recuperación o análisis que se quisiese hacer sobre los datos almacenados en la base de datos tendría a su vez problemas a la hora de interpretar tal variedad. El uso de vocabularios controlados (CVs, *Controlled Vocabularies*) y bio-ontologías (ver sección 0) públicas pareció ser la solución. Actualmente los cuatro módulos MIAPE implementados tienen asociados un gran número de términos provenientes de diferentes ontologías, de tal manera que hay secciones concretas de los documentos en las que el usuario deberá seleccionar un valor de un desplegable compuesto por vocabularios controlados (Figura 40). Esto permite una validación semántica de la información que se introduce en la herramienta, y elimina la heterogeneidad a la hora de describir los conceptos requeridos por las directrices MIAPE.

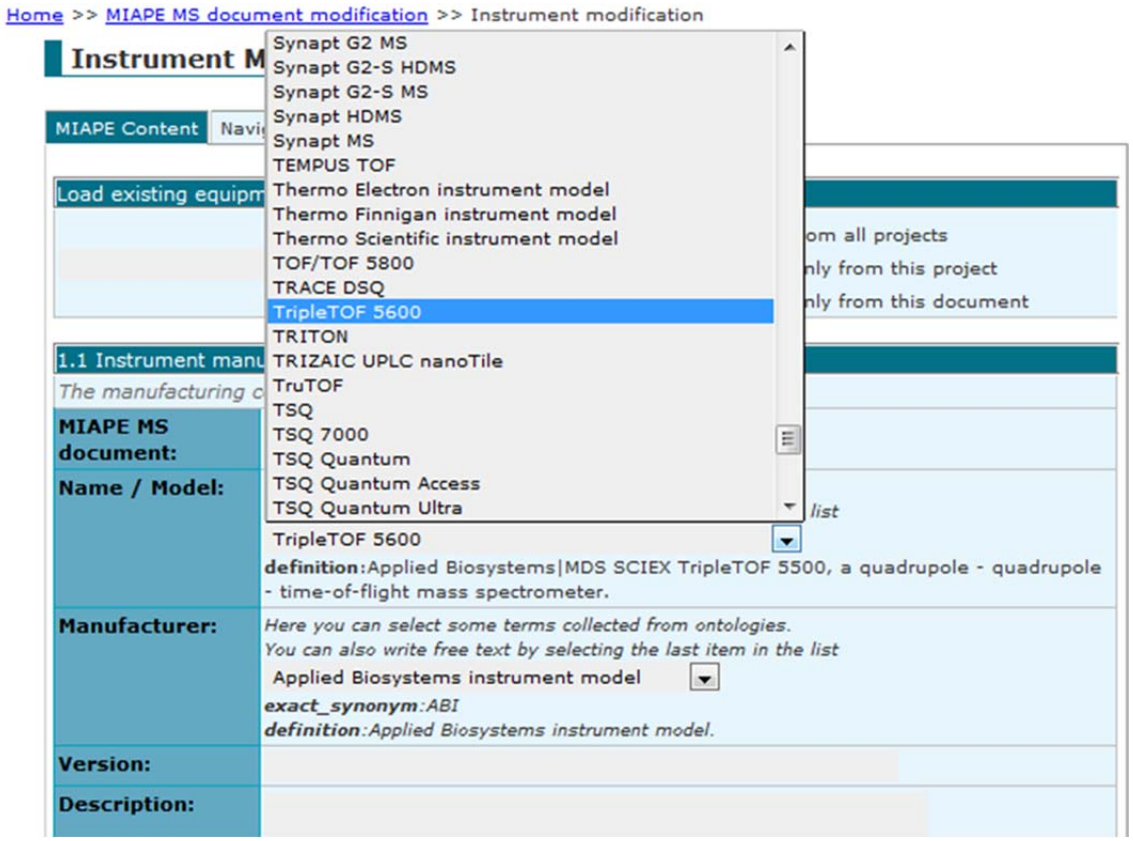


Figura 40. Desplegables de vocabularios controlados en la herramienta generadora de documentos MIAPE: Un conjunto de vocabularios controlados se asocian a cada sección MIAPE. En este caso, en el módulo de *Mass Spectrometry*, se muestra la sección sobre el espectrómetro, donde tanto el nombre del espectrómetro (desplegado) como el fabricante (*manufacturer*) se pueden especificar con listas desplegables.

La relación entre las diferentes secciones de los cuatro módulos MIAPE y los numerosos términos de las ontologías disponibles en el servicio OLS (*Ontology Lookup Service*) (Cote, Jones et al. 2006) ha sido un trabajo clave tanto para la herramienta generadora de documentos MIAPE como en las subsiguientes herramientas, ya que permite la interpretación programática y por tanto automática de la información. Esto será por tanto la base de la automatización de todos los procesos de validación e informe de los datos de acuerdo a las directrices MIAPE. Una tabla con el mapeo de las secciones MIAPE y los vocabularios controlados se mostrará en la sección 4.2.2, referente a la validación de los ficheros estándares mzML y mzIdentML.

La herramienta generadora de documentos MIAPE ha sido utilizada también en el contexto del HUPO-PSI en diferentes grupos de trabajo a la hora de redefinir nuevas versiones de algunos módulos MIAPE ya existentes, incluyendo documentos MIAPE generados por la herramienta como parte de los ejemplos requeridos para la aprobación de los módulos en el proceso de revisión del PSI (Vizcaino, Martens et al. 2007). Además, la herramienta se ha ido adaptando a los cambios establecidos tanto en las diferentes versiones que ha habido de cada

módulo MIAPE, como a los nuevos vocabularios controlados definidos en las ontologías, por lo que siempre ha sido la referencia para las últimas versiones de las especificaciones. Pese a los cambios, la información almacenada en la base de datos siempre ha mantenido su integridad.

Como veremos más adelante, subsiguientes desarrollos se han centrado en la automatización de la extracción y almacenamiento en la base de datos de la información MIAPE, con lo que el número de documentos almacenados ha crecido exponencialmente gracias a ello. Sin embargo y pese a que la herramienta generadora de documentos MIAPE requiere siempre de una intervención manual (mayor o menor) por parte del usuario, en el momento de su publicación (Martinez-Bartolome, Medina-Aunon et al. 2010), tras poco más de un año de funcionamiento, ya existían más de setecientos documentos MIAPE almacenados en la base de datos, la mayoría de ellos pertenecientes a experimentos reales y sólo algunos correspondientes a plantillas o pruebas de algún usuario.

Por tanto, la herramienta generadora de documentos MIAPE no sólo sirve como herramienta de ayuda para generar documentos MIAPE sino que además constituye el primer repositorio de documentos MIAPE, es decir, la primera base de datos con información de experimentos completos de Proteómica, basados en geles y de identificación por espectrometría de masas. Además permite la comparación de distintos experimentos para así, por ejemplo, identificar las características de un protocolo que permiten obtener mayor rendimiento que otros.

4.2.2. Desarrollo de herramientas de validación semántica y MIAPE de estándares de representación de datos

Como se puede ver en la Tabla 4, existe una relación directa entre los diferentes formatos estándares de representación de datos y los módulos MIAPE. Cada formato XML debería contener la información requerida en el módulo MIAPE que le corresponde, para lo cual se hace necesario disponer de herramientas que lean estos ficheros y comprueben si falta alguna información requerida por las directrices. Sin embargo, las únicas herramientas disponibles para validar los ficheros estándares son las descritas en la sección 0, que únicamente sirven para validar archivos mzML, mzIdentML o PRIDE XML de manera semántica, no teniendo en cuenta todos los detalles descritos en las directrices MIAPE. Es por ello por lo que quisimos ampliar la funcionalidad de las herramientas de validación existentes para comprobar el seguimiento de las directrices MIAPE correspondientes.

4.2.2.1. Desarrollo de las herramientas de validación semánticas de los ficheros mzML y mzIdentML

Los ficheros mzML (Deutsch 2008, Martens, Chambers et al. 2011) y mzIdentML (Eisenacher 2011, Jones, Eisenacher et al. 2012) son cada vez más utilizados y cada vez son más las herramientas bioinformáticas que los generan, o utilizan. Sin embargo, y pese a que los estándares generados por el PSI se publican con una extensa documentación, en la que también se incluye una correspondencia entre el correspondiente módulo MIAPE y los diferentes elementos que los forman, no siempre los ficheros generados son válidos semánticamente, y más frecuentemente, no suelen seguir las directrices MIAPE conteniendo la información mínima para describir la parte correspondiente del experimento que representan.

Así pues, partiendo del código en Java utilizado por el grupo de PRIDE del EMBL-EBI (*Hinxton-Cambridge*), se ampliaron las herramientas de validación semántica existentes para los ficheros mzML y mzIdentML. Estas herramientas están basadas en la librería de validación semántica del PSI (Montecchi-Palazzi, Kerrien et al. 2009), cuyo funcionamiento consiste en definir una serie de reglas de validación, las cuales se aplicarán luego a los diferentes elementos del XML del fichero a validar. Estas reglas pueden ser de dos tipos:

- **Reglas de mapeo de vocabularios controlados** (*cvMappingRule*): como se comentó en la introducción estas reglas definen qué conjunto de términos está permitido en cada uno de los elementos del fichero XML y con qué severidad (sugerencia->MAY, conveniencia->SHOULD u obligatoriedad->MUST). Por ejemplo, la regla:

```
<CvMappingRule
        id="SoftwareName_rule"
    cvElementPath="/MzIdentML/AnalysisSoftwareList/AnalysisSoftware/softwareName/cvParam/
    @accession"
        requirementLevel="MUST"
        scopePath = "/MzIdentML/AnalysisSoftwareList/AnalysisSoftware/softwareName"
    cvTermsCombinationLogic="OR">
            <CvTerm termAccession="MS:1001456"
                    useTermName="false"
                    useTerm="false"
                    termName="analysis software"
                    isRepeatable="true"
                    allowChildren="true"
                    cvIdentifierRef="MS"/>
</CvMappingRule>
```

  significa que dentro del elemento *softwareName*, cuyo XPath (ruta en el fichero XML) es "*/MzIdentML/AnalysisSoftwareList/AnalysisSoftware/softwareName*" tiene que haber obligatoriamente (*MUST*) uno o más elementos cvParam (*isRepeatable="true"*), conteniendo términos hijos (*allowChildren="true"*) del término "analysis *software*" (*MS:1001456*) de la ontología *MS* y no el mismo término (*useTerm="false"*).

- **Reglas de objeto** (*objectRule*): son reglas programáticas sobre un tipo de objeto que representa la sección a validar. Estas reglas permiten mayor flexibilidad, pudiéndose validar cualquier cosa que se pueda implementar programáticamente. Ejemplo:

```
Public Collection<ValidatorMessage> check (SpectrumIdentificationProtocol protocol) throws
ValidatorException {
        List<ValidatorMessage> messages = new ArrayList<ValidatorMessage>();
        if (protocol.getParentTolerance() == null) {
                messages.add (new ValidatorMessage("There is not a parent tolerance!",
MessageLevel.ERROR, specIdenProtocolConContext, this));
        }
        return messages;
}
```

Esta regla, valida los objetos del elemento "*SpectrumIdentificationProtocol"* comprobando si el subelemento "*ParentTolerance*" está presente o no, y si no lo está, genera un mensaje de error.

El primer paso por tanto para construir los validadores MIAPE fue el mapeo entre la información requerida en las directrices MIAPE y los elementos en los ficheros XML en los que debería ir dicha información, así como los vocabularios controlados que permiten anotarla en el fichero. Para ello, se desgranó lo máximo posible el lenguaje natural con el que las directrices MIAPE MS y MSI estaban definidas, y cada unidad de información requerida se mapeó con el estándar y con la ontología PSI-MS. Como resultado, obtuvimos las tablas en el material suplementario para el mzML (Tabla 9) y para el mzIdentML (Tabla 10).

Gracias a este trabajo, numerosos términos fueron añadidos a la ontología PSI-MS para definir información no existente hasta la fecha, y necesaria para anotar información MIAPE, como "*sprayer*", "*source interface*", "*acquisition parameters*", etc…

Todas estas relaciones entre elementos XML, vocabularios controlados e información MIAPE, se tradujeron en lo posible a reglas de mapeo de vocabularios controlados (*cvMappingRule*) o reglas de objeto (*objectRule*), creándose así un nuevo conjunto de reglas. Los ficheros de mapeo de CVs están accesibles en: https://psidev.svn.sourceforge.net/svnroot/psidev/psi/mzml/validator/miape-ms-rules.xml para el mzML, y http://psi-pi.googlecode.com/svn/trunk/validator/trunk/miape-msi-rules.xml para el mzIdentML.

Pese a todo este esfuerzo, ciertos detalles especificados de las directrices MIAPE seguían sin poderse validar con este tipo de reglas, y es que muchas veces, dependiendo del tipo de experimento, o simplemente, de un único aspecto de éste, tiene sentido esperar ciertas anotaciones u otras, o lo que es lo mismo, tiene sentido aplicar un conjunto de reglas u otras.

Para ello, se implementó un módulo por encima de los validadores que permitía una **aplicación condicional de las reglas**. La condición podría ser:

- la entrada que el usuario hiciese por medio de la interfaz gráfica o
- el resultado de la aplicación de una regla en concreto.

Para el primer caso, se modificaron las interfaces gráficas para permitir al usuario, además de elegir entre la validación semántica o MIAPE, y la severidad de la validación (*Info*, *Warn o Error*), introducir ciertos aspectos del experimento (Figura 41 y Figura 42).

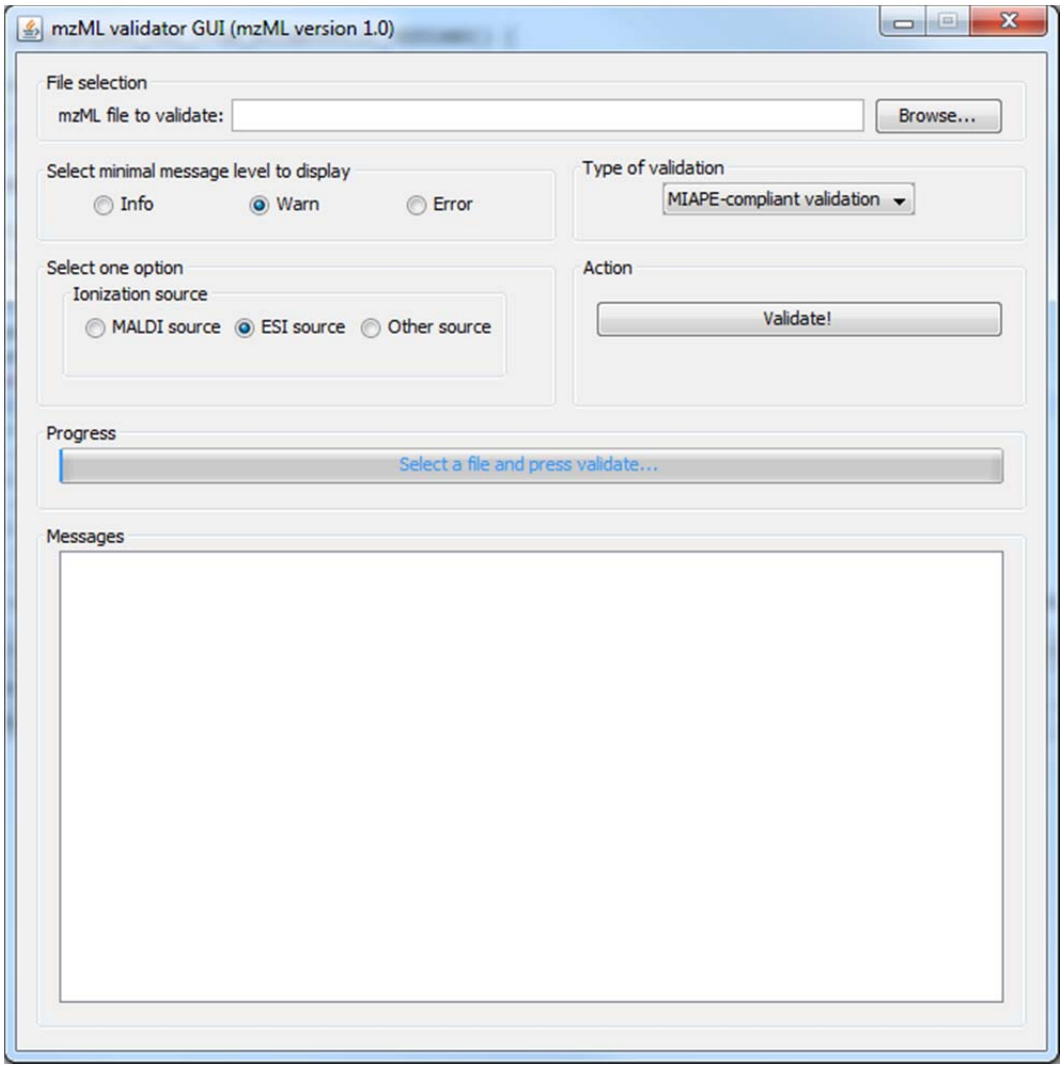


Figura 41. Interfaz gráfica del validador semántico y MIAPE de ficheros mzML: En la interfaz el usuario debe seleccionar el fichero mzML, seleccionar el nivel de severidad de los resultados, seleccionar el tipo de validación (semántica o MIAPE) y seleccionar el tipo de ionización que se ha utilizado: MALDI, ESI u otra.

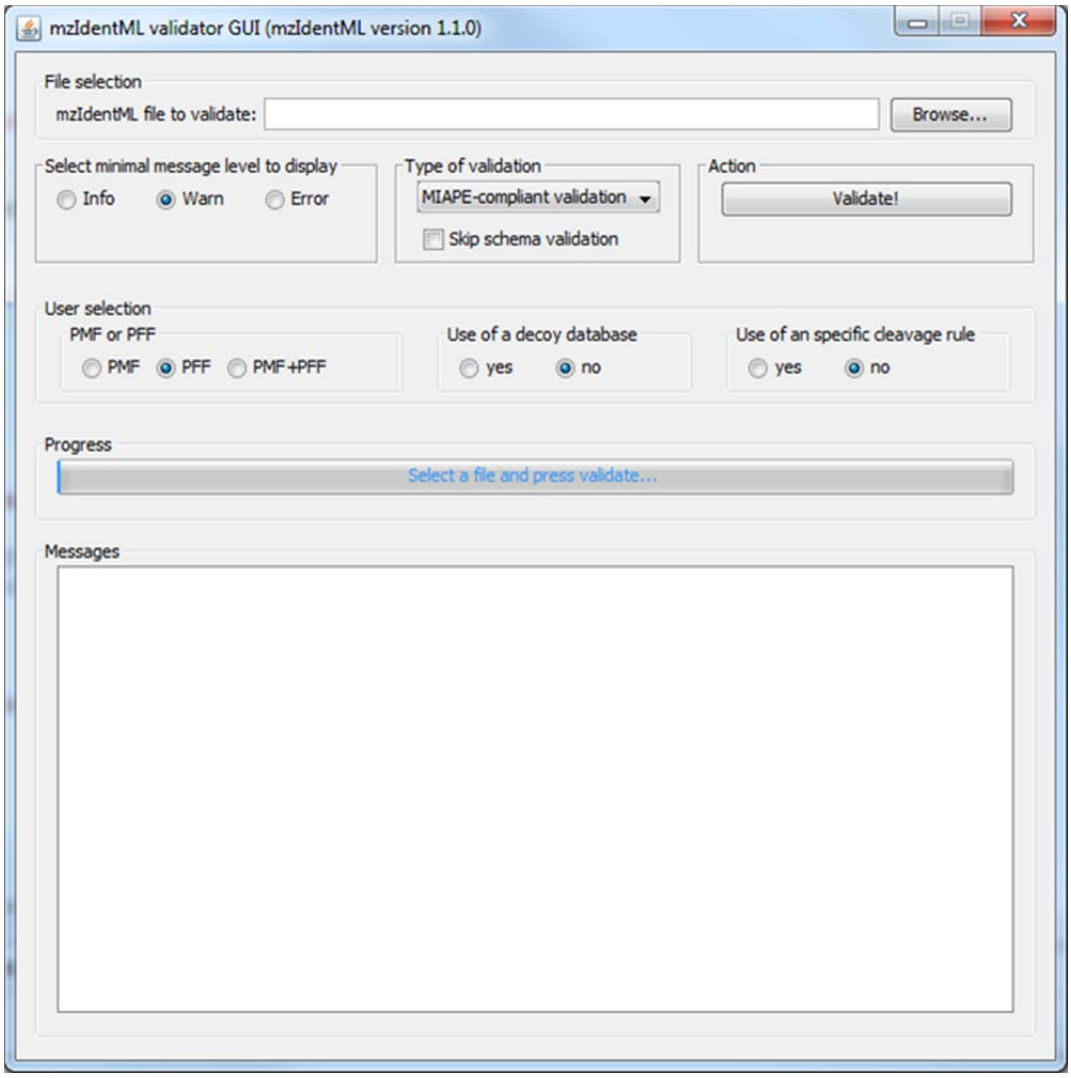


Figura 42. Interfaz gráfica del validador semántico y MIAPE de ficheros mzIdentML: En la interfaz el usuario debe seleccionar el fichero mzIdentML, seleccionar el nivel de severidad de los resultados, seleccionar el tipo de validación (semántica o MIAPE) y seleccionar tres opciones: si se ha realizado una búsqueda sobre datos de huella peptídica (PMF) o de fragmentación (PFF); si se ha utilizado una base de datos señuelo (decoy); y si se ha utilizado una regla propia de corte enzimático (*cleavage rule*).

En el caso de las reglas condicionadas al resultado de otras, se definen qué reglas van a servir como condición, y dependiendo de si se cumplen o no, se define qué otras reglas no se van a aplicar, o más bien, se van a ignorar.

Veamos algunos ejemplos:

Como condiciones de opción de usuario:

- Para el mzML: "*si el usuario dice que la fuente de ionización es MALDI, no se aplicarán las reglas que comprueban el tipo de fuente ESI, el espray y la interfaz*".

- Para el mzIdentML: "*si el usuario dice que se ha realizado una búsqueda de huella peptídica, no se comprobará el parámetro de búsqueda de tolerancia de fragmento*".

Como condiciones de regla:

- Para el mzML: "*si la regla que comprueba si se ha especificado un fichero de parámetros de adquisición no devuelve errores, esto es, que efectivamente se ha especificado ese fichero, no se aplicará la regla que comprueba que los parámetros de adquisición se han especificado explícitamente en el elemento del software de adquisición*".
- Para el mzML: "*si la regla que comprueba si se ha especificado un analizador tipo TOF devuelve algún error, es decir, el analizador es otro distinto que el TOF, entonces no se aplicará la regla que comprueba si se ha especificado el estado del reflectrón del analizador*".

En realidad, el validador aplica todas las reglas internamente, independientemente de las condiciones de opciones de usuario o de regla. Sin embargo antes de mostrar al usuario los mensajes de validación resultantes, éstos se filtran en función de las opciones seleccionadas por el usuario en la interfaz y en función de los resultados de las reglas que estén implicadas en alguna condición de regla. Así pues, el esquema de funcionamiento de la ampliación de los validadores se resume en la Figura 43.

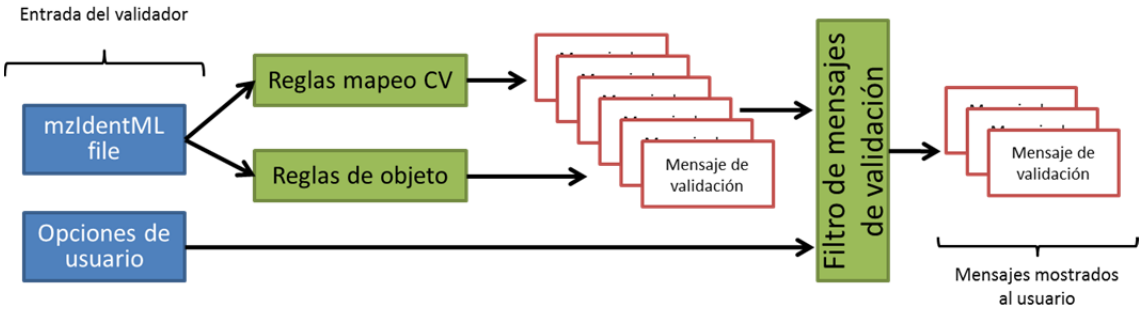


Figura 43. Esquema de funcionamiento de los validadores de ficheros estándares: El usuario introduce el fichero, en este caso mzIdentML, y selecciona las opciones correspondientes en la interfaz de usuario (entrada del validador). Luego, todas las reglas de mapeo de CV y de objeto son ejecutadas sobre el fichero en cuestión. Estas reglas producirán un conjunto de mensajes de validación correspondientes a las reglas que no se hayan cumplido. Luego, estos mensajes pasan por un filtro en el cual se tienen en cuenta las dependencias entre reglas y las opciones seleccionadas por el usuario. Finalmente se muestran al usuario los mensajes de validación que pasen el filtro (mensajes mostrados al usuario).

Tanto las condiciones de usuario como las condiciones de reglas, se definen en un fichero de configuración en XML muy sencillo, que incluso podría ser modificado por un usuario avanzado que quisiese definir sus propias reglas y condicionarlas a su manera.

En el caso del validador de ficheros mzIdentML, se incluyó dentro de una publicación donde se describían un conjunto de herramientas relacionadas con dicho estándar: además del validador, un visualizador *ProteoIDViewer* y una librería *mzidLibrary* Java para realizar diferentes operaciones comunes con los mzIdentML, además del validador (Ghali, Krishna et al. 2013).

4.2.2.2. Desarrollo de una herramienta para validar los documentos MIAPE de experimentos basados en geles

Como hemos dicho, las únicas herramientas que realizaban algún tipo de validación sobre los datos contenidos en los ficheros estándares trabajaban en todo caso con la parte de espectrometría de masas y su identificación en bases de datos, pero en ningún caso con datos provenientes de experimentos de separación de proteínas basada en geles, esto es, con el estándar gelML (Gibson, Hoogland et al. 2010).

Por ello, y dado que teníamos ya un repositorio con datos relativos a experimentos de geles, esto es, los documentos MIAPE GE y GI, desarrollamos un procedimiento, incluido en el ProteoRed MIAPE Web Toolkit (PMWTK) (Medina-Aunon, Martinez-Bartolome et al. 2011), para generar un fichero gelML con una nueva herramienta, el *gelML exporter*. Esta herramienta, desarrollada principalmente por Alberto Medina (Centro Nacional de Biotecnología - CSIC, Madrid), genera un fichero gelML a partir de la información almacenada en un documento MIAPE GE de la base de datos. Sin embargo, dado que la introducción de los datos del documento MIAPE GE es manual o semi-automática en la herramienta generadora de MIAPEs, se hacía imprescindible un paso previo al *gelML exporter* para comprobar que el documento MIAPE GE tuviese la información mínima necesaria para construir luego el fichero gelML y que esa información estuviese correctamente anotada y relacionada. Así pues, se desarrolló un validador para la información MIAPE Gel Electrophoresis previo a la generación del fichero gelML.

Esta herramienta de validación se integró complemente en la interfaz web de la herramienta generadora de documentos MIAPE, estando disponible un botón para cada documento MIAPE GE almacenado (Figura 44 A). Pinchando en este botón, aparecerá la página de validación (Figura 44 B), donde el usuario, tras pulsar en otro botón, recibirá los mensajes de error correspondientes en su caso (Figura 44 C), o un mensaje de confirmación de que el documento está preparado para ser transformado en un fichero gelML (Figura 44 D). En los mensajes de validación se muestran dos tipos de severidad, error o advertencia, según sean aspectos insalvables para la generación del fichero gelML o sean aspectos opcionales sin los cuales el fichero gelML resultante sería igualmente válido.

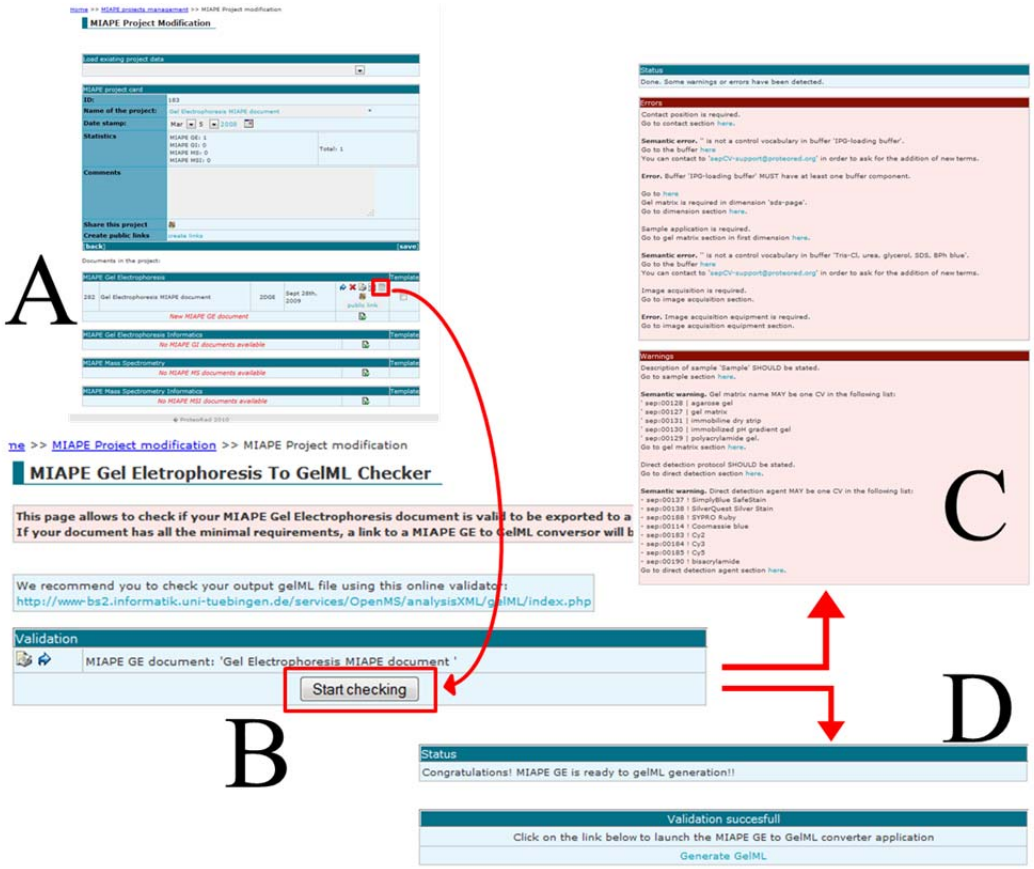


Figura 44. Funcionamiento del validador de documento MIAPE GE: (A) Pinchando en el botón "*export to gelML*" (icono de imagen de gel) que se encuentra junto con las opciones de editar, borrar, etc... del documento MIAPE GE, nos lleva a la página de validación. En la página de validación (B), pinchando en el botón "*Start checking*" el documento MIAPE GE será validado. El resultado de la validación se muestra luego al usuario, si el documento tiene algún error o incoherencia (C), en cuyo caso se proporciona un link a la sección errónea para solucionarla, o si está totalmente correcto (D).

Las reglas de validación, en este caso, están implementadas directamente en el código, sin ningún fichero de configuración que las defina, debido a la dificultad que entrañaría implementarlo de manera configurable en la tecnología ASP. Son por tanto análogas a las reglas de objeto descritas anteriormente. Algunas de las reglas implementadas por este validador son:

- *Para cada disolución tamponada (búfer) comprueba que el "tipo de búfer" asociado está anotado con un vocabulario controlado hijo del término "buffer solution" (sep:00118).*
- *Si en la primera sección del documento se define que el experimento es un experimento de geles de dos dimensiones ("two dimensional gel electrophoresis" - sep:00155) o ("difference gel electrophoresis" - sep:00180), deberá haber definidas dos secciones "Dimension". Si se define como un gel de una sola dimensión ("one dimensional gel electrophoresis" - sep:00150) sólo podrá haber una sección "Dimension".*

- *Todas los datos asociados a las substancias definidas (masa, volumen, concentración, etc…) deberán ir acompañadas de sus correspondientes unidades, además, por supuesto de ser números positivos.*
- *Si se especifica en qué carril se aplica cada muestra en el gel, debe asociarse cada carril a un número creciente, y cada uno debe contener una referencia a una muestra definida en la sección correspondiente.*

El obtener un documento MIAPE GE validado, nos asegura que dicho documento está bien formado, que la información que contiene tiene sentido semántico y que no falta ninguna información crucial para la descripción del experimento. Por tanto, el siguiente paso en el que se genera un fichero gelML estándar se convierte en una tarea trivial usando la herramienta *gelML exporter*.

El validador de documentos MIAPE GE también se puede utilizar durante la creación de un documento MIAPE describiendo un experimento basado en separación por geles, simplemente para comprobar que la información que se ha introducido es correcta, lo cual ayuda enormemente a completarlo.

4.2.3. Desarrollo de la librería para la extracción y manejo de la información MIAPE

Como base para la automatización de la extracción, almacenamiento y análisis de la información MIAPE, se desarrolló una librería, la llamada *ProteoRed Java MIAPE API*, cuyo código fuente está disponible en google code: http://code.google.com/p/proteored-java-miape-api/.

La librería fue diseñada de manera totalmente modular, conteniendo principalmente 3 partes diferenciadas (Figura 45):

I. El módulo que define el modelo de objetos o de datos: define la jerarquía de información MIAPE de los cuatro tipos de documentos MIAPE soportados (MIAPE GE, GI, MS y MSI). Para cada sección MIAPE, se construye una clase conteniendo la información de dicha sección en forma de atributos de la clase y las subsecciones en forma de colecciones de objetos de la clase de la subsección.

II. El módulo XML, que implementa las funcionalidades de conversión y extracción de información MIAPE desde y hacia ficheros XML.

III. El módulo de factorías de objeto, que permite la creación programática de objetos de las clases MIAPEs implementados de forma totalmente intuitiva.

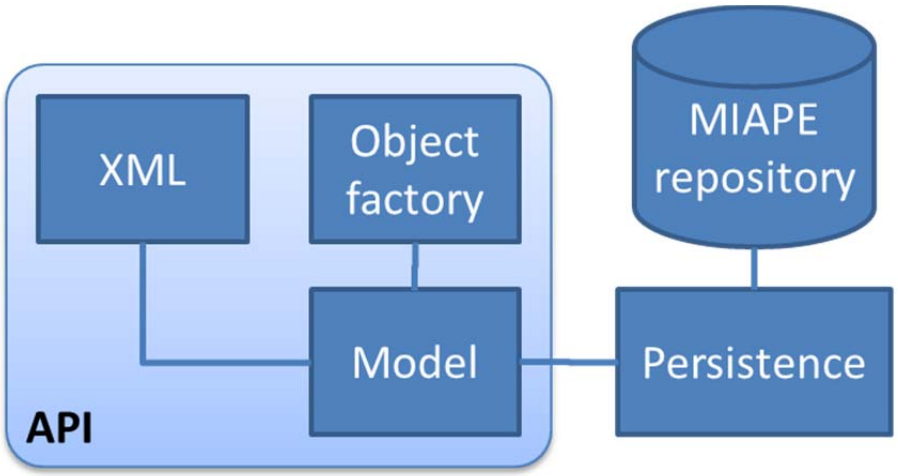


Figura 45. Modularidad de la librería MIAPE: Los módulos que componen la librería MIAPE son: el módulo XML, para la conversión y extracción de información MIAPE de ficheros XML, el módulo de factoría de objetos, para la creación programática de documentos MIAPE y el módulo del modelo de datos, que define la información que contiene cada documento MIAPE. El módulo de persistencia, definido en la librería, aunque no implementado en ella, permite la interacción con un sistema de persistencia (como una base de datos).

El módulo XML se diseñó de forma muy funcional. En primer lugar se diseñaron 4 esquemas XML para permitir guardar los documentos MIAPE en ficheros XML (http://proteo.cnb.csic.es/miape-api/schemas/), MIAPE GE XML, GI XML, MS XML y MSI XML. El módulo XML permite la conversión bidireccional entre el modelo de objetos y dichos ficheros. Además, permite la extracción de la información MIAPE de diferentes ficheros XML, como son los estándares PRIDE XML, mzML, mzIdentML y mzIdentML, u otros ficheros como el fichero de salida del buscador X!Tandem. Una vez extraída la información MIAPE, se lleva al modelo de objetos, donde se accede a los diferentes datos MIAPE. A su vez, el módulo XML permite la conversión del modelo de datos en diferentes ficheros XML, en este caso, en un fichero PRIDE XML o en los 4 tipos de ficheros MIAPE XML (Figura 46). Así pues, gracias al módulo XML de la librería podremos crear ficheros PRIDE XML que contienen toda la información proveniente de nuestros documentos MIAPE MS y MSI, siendo la primera vez que se proporciona una manera de generar ficheros PRIDE XML que siguen las directrices MIAPE, es decir, que describen perfectamente un experimento de identificación de péptidos y proteínas por espectrometría de masas con todos sus metadatos.

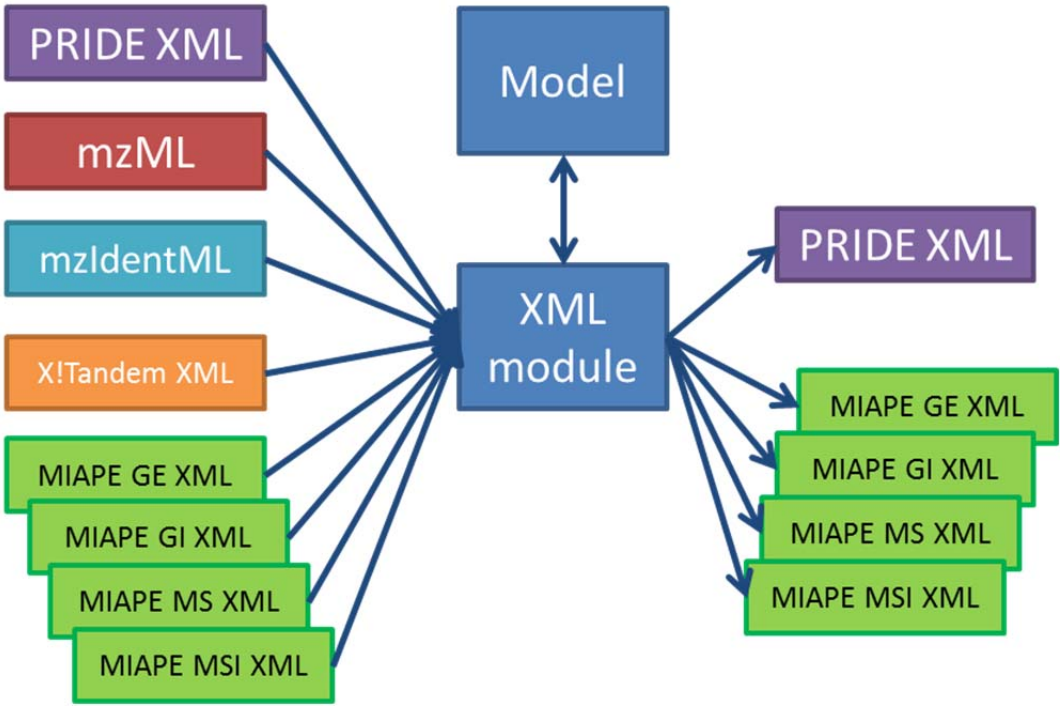


Figura 46. Esquema de funcionamiento del módulo XML de la librería de MIAPEs: El módulo XML de la librería MIAPEs es capaz de extraer la información MIAPE de diferentes ficheros en XML, entre los que están los estándares PRIDE XML, mzML, mzIdentML y mzIdentML, otros ficheros como el fichero de salida del buscador X!Tandem, y los ficheros XML definidos internamente por nosotros para representar exclusivamente la información MIAPE de los 4 módulos en un fichero XML. La información MIAPE es convertida en el modelo de objetos y puede ser a su vez convertida en un fichero PRIDE XML o en los 4 tipos de ficheros XML MIAPE.

El módulo de persistencia (Figura 45) permite el almacenamiento de la información MIAPE en un medio de persistencia, por ejemplo, una base de datos. Por razones de seguridad este módulo no está implementado sobre el repositorio de MIAPEs de ProteoRed en la librería pública, ya que entonces se daría acceso a todos los datos allí almacenados sin ningún control.

4.2.4. Desarrollo de un acceso programático al repositorio de experimentos proteómicos

Así pues, para el acceso programático pero controlado de la información MIAPE almacenada en el repositorio, se implementaron también unos servicios web, que permiten de manera remota acceder y almacenar de manera segura documentos MIAPE, manteniendo el control de acceso y los distintos ámbitos de los usuarios (Figura 47). Los **servicios web** (*web-services*), proporcionan una serie de métodos remotos a los que puede acceder cualquier programa externo, independientemente de la plataforma o lenguaje en el que esté implementado.

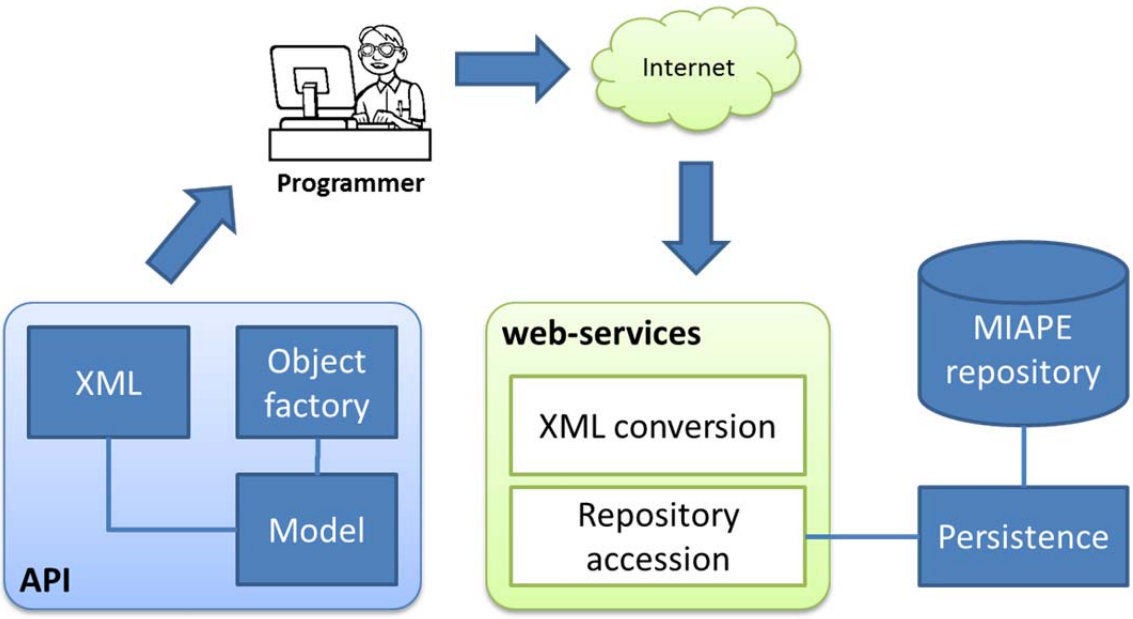


Figura 47. Esquema general de la librería MIAPE y el acceso remoto al repositorio MIAPE de ProteoRed por medio de los servicios web: Un programador podrá utilizar la librería MIAPE utilizando los tres módulos en los que está compuesta, y para acceder al repositorio de MIAPEs de ProteoRed podrá utilizar el acceso remoto que le proporcionan los servicios web. Además, estos servicios web permiten la extracción y almacenamiento de información MIAPE en el repositorio a partir de ficheros XML estándares.

Así pues, este acceso remoto puede permitir el desarrollo de nuevas herramientas, incluso externas, que trabajen con la información de nuestro repositorio o que sirvan como nuevos puntos de acceso para introducir datos en él.

4.2.5. Desarrollo de una herramienta para proporcionar un flujo completo de integración, análisis e informe de datos siguiendo las directrices MIAPE

Gracias a la plataforma de desarrollo que proporcionan el tándem formado por la librería Java de MIAPEs y los servicios web que permiten acceder al repositorio de MIAPEs de ProteoRed, desarrollamos una herramienta bioinformática con un primer objetivo: la automatización de la creación de los documentos MIAPE y los ficheros PRIDE XML conformes a las directrices MIAPE. Una vez cumplido dicho objetivo la herramienta fue paulatinamente ampliada hasta permitir un flujo completo de manejo, almacenamiento, análisis y reporte de datos.

Este flujo de trabajo ha sido incorporado en el día a día de nuestro laboratorio para el análisis de identificaciones a gran escala, y lo que es más importante, ha sido el flujo de trabajo

que ha seguido el grupo de trabajo de identificación a gran escala de la iniciativa española del Proyecto del Proteoma Humano (Sp-HPP – *Spanish Human Proteome Project*), en la que se pretende caracterizar todos los productos proteicos derivados de la expresión de los genes del cromosoma 16 humano (Segura, Medina-Aunon et al. 2013, Segura, Medina-Aunon et al. 2013).

El flujo de trabajo que finalmente permite la herramienta está compuesto por diferentes fases (Figura 48):

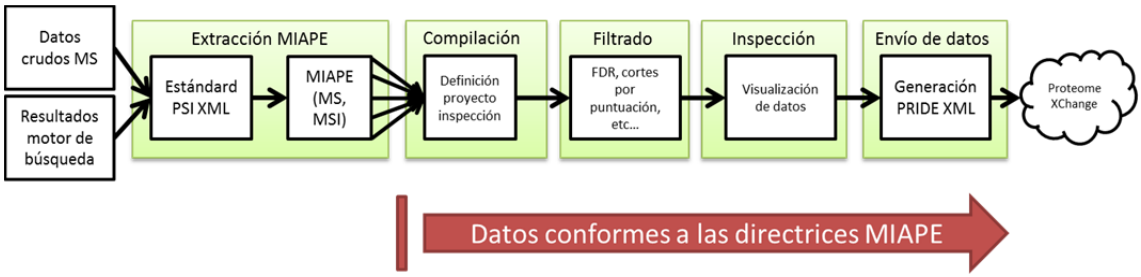


Figura 48. Flujo completo de datos de la herramienta MIAPE Extractor: Tanto los datos crudos provenientes del espectrómetro de masas como los resultados de los buscadores son introducidos en la herramienta, tras su conversión a los estándares PSI en XML correspondientes. Luego, se extrae la información MIAPE MS y/o MSI, la cual es también almacenada en el repositorio de documentos MIAPE de ProteoRed. Después, el usuario define con qué datos provenientes de MIAPEs MSI quiere trabajar y construye un proyecto de inspección de datos. Una vez definido el proyecto, el usuario puede aplicar una serie de filtros a los datos seleccionados. Además, la herramienta permite inspeccionar dichos datos por medio de numerosas gráficas. Finalmente, la herramienta permite la creación de todo lo necesario para un envío de datos a ProteomeXchange.

1) Extracción de la información MIAPE

Una versión preliminar de la herramienta fue descrita en la publicación (Medina-Aunon, Martinez-Bartolome et al. 2011). En dicha versión, la herramienta era capaz de extraer información MIAPE MSI de los ficheros estándares mzIdentML (v1.0.0) y PRIDE XML. En posteriores versiones de la herramienta, desarrollándose en paralelo a la librería de MIAPEs, la extracción de información MIAPE es posible desde ficheros mzIdentML versión 1.0.0 y 1.1.0, ficheros PRIDE XML, y ficheros XML de resultados del buscador X!Tandem, todos ellos para extraer la información MIAPE MSI, y ficheros mzML para extraer la información MIAPE MS.

El control de usuarios es el mismo que en la herramienta online de MIAPEs. Tras iniciar una sesión en la herramienta el usuario (proporcionando el nombre de usuario y la contraseña) tiene que seleccionar el o los ficheros de los que se quiere extraer la información MIAPE y el proyecto en el que se quieren crear. La herramienta utilizará los servicios web desplegados en el servidor para subir automáticamente los ficheros, donde serán procesados para almacenar los

documentos MIAPE en la base de datos. El servicio web, tras terminar su trabajo, devolverá a la herramienta el identificador de los documentos creados en la base de datos.

Opcionalmente, el usuario podrá seleccionar que el procesado de los ficheros sea local, con lo que no se subirá al servidor más que la información MIAPE extraída localmente. Esto permite que los usuarios reticentes a subir sus ficheros a un servidor externo puedan utilizar la herramienta igualmente. En próximas versiones se dará la posibilidad de que tras extraer la información MIAPE localmente, ésta se pueda procesar en los pasos posteriores del flujo de trabajo sin necesidad de pasar por el repositorio de MIAPEs de ProteoRed.

Cada vez son más las herramientas que permiten generar los estándares de representación de datos de espectrometría de masas, convirtiendo los datos crudos del espectrómetro al formato mzML o exportando los resultados de los motores de búsqueda al formato mzIdentML. Pese a ello, los ficheros resultantes no siempre contienen todos los datos requeridos por las directrices MIAPE MS y MSI. En especial, esto es bastante frecuente en el caso de los ficheros mzML, ya que los conversores de los diferentes datos crudos propietarios extraen muy pocos metadatos acerca del instrumento y los parámetros de adquisición de espectros.

Para ilustrar ese problema, se inspeccionó la información MIAPE contenida en diferentes ficheros mzML creados a partir de diferentes datos crudos y usando diferentes herramientas transformadoras y se construyó la Tabla 6. Como se puede observar, la mayoría de los metadatos de la instrumentación utilizada en la adquisición de espectros no están disponibles en los ficheros mzML resultantes (celdas en rojo). Por tanto, los documentos MIAPE MS creados a partir de esos ficheros necesariamente estarán incompletos. Para completarlos, la herramienta online generadora de documentos MIAPE permite editar los documentos almacenados, para así añadir la información faltante. Una vez actualizado, el documento resultante podría utilizarse de nuevo para generar el fichero PRIDE XML completo y cumpliendo las directrices MIAPE. Sin embargo, se añadió un editor de metadatos de espectrometría de masas en la propia herramienta MIAPE Extractor, para no perder la automatización del proceso y así generar directamente documentos MIAPE MS completos gracias a los ficheros mzML más los metadatos incluidos en la herramienta (Figura 49).

			Thermo LTQ Orbitrap XL	Waters Synapt	ABI Qstar XL / 5600 Ttof	ABI 4800 MALDI TOF/TOF	Bruker HCT Ultra	Thermo LTQ Orbitrap XL	Bruker HCT Ultra	ABI 4800 MALDI TOF/TOF	ABI 5600 Ttof
			msConvert 3.0.3643					Proteome Discoverer 1.3.0.0	Compass Xport 3.0.5	ABSciex MS Data Converter vBeta 1.1	
MIAPE MS section			Thermo RAW	Waters RAW	AB wiff	AB t2d	Bruker / Agilent YEP	Thermo RAW	Bruker / Agilent YEP	AB t2d	AB wiff
General features		Responsible person									
Instrument		Name									
		Manufacturer									
		Version - S/N									
Ion source	ESI	Supply type									
		Interface									
		Sprayer									
	MALDI	Plate	N.A.	N.A.	N.A.		N.A.	N.A.	N.A.		N.A.
		Matrix	N.A.	N.A.	N.A.		N.A.	N.A.	N.A.		N.A.
		PSD/LID/ISD	N.A.	N.A.	N.A.		N.A.	N.A.	N.A.		N.A.
		Laser	N.A.	N.A.	N.A.		N.A.	N.A.	N.A.		N.A.
Post-source component		Analyser									
		Activation / dissociation									
		Gas									
		Activation type									
Data Acquisition		Softw. name									
		Softw. version									
		Parameters									
Data Analysis (data conversion)		Softw. name									
		Softw. version									
		Parameters									
Resulting Data											

Tabla 6. Estudio de la información MIAPE contenida en diferentes ficheros mzML generados por diferentes software

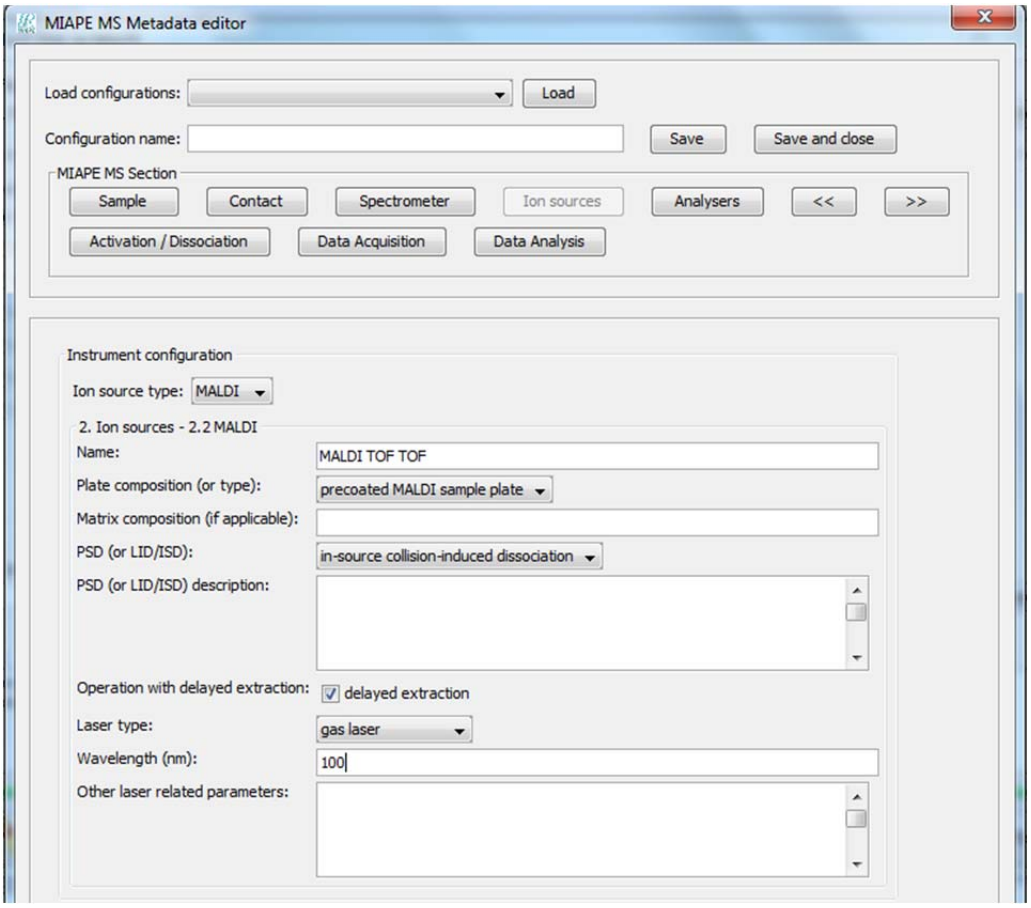


Figura 49. Editor de metadatos MIAPE MS: La herramienta MIAPE Extractor proporciona la posibilidad de introducir los metadatos no extraídos de los ficheros de entrada requeridos por el módulo MIAPE MS. En este caso, se muestran los formularios para introducir los datos de la fuente de ionización MALDI. Estos metadatos se pueden guardar como plantillas y así poder utilizarlos en sucesivas ocasiones.

Para una completa automatización de la extracción de información MIAPE se implementó un sistema de colas de trabajos de extracción. Dichos trabajos se definen en un fichero de texto muy sencillo donde se le indica qué ficheros se quieren utilizar y en qué proyecto se quieren asociar los MIAPEs. Una vez cargado el fichero, la herramienta mostrará el progreso de cada una de las extracciones, que se ejecutarán una detrás de otra (Figura 50).

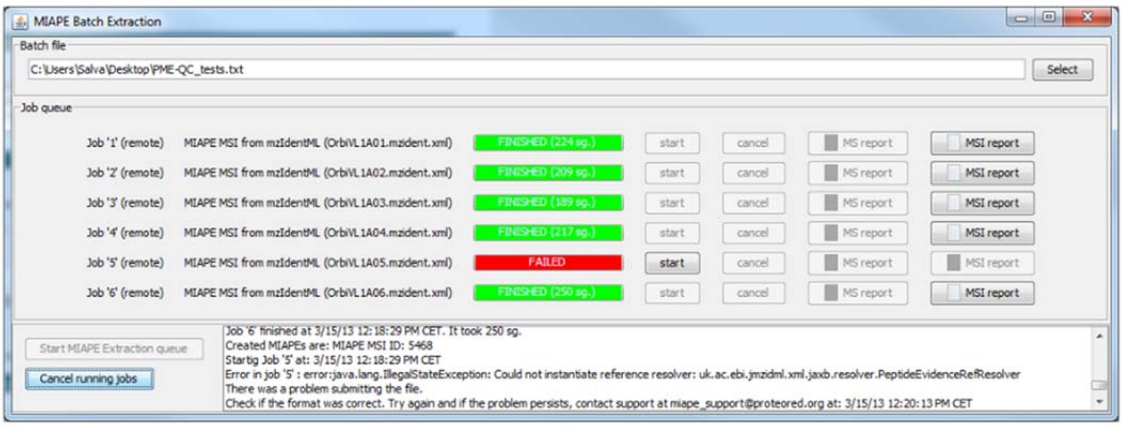


Figura 50. Interfaz de la extracción MIAPE en cola: Una vez cargado el fichero de texto indicando los trabajos de extracción, éstos se ejecutan uno detrás de otro. La herramienta mostrará en verde los documentos creados correctamente y en rojo los que han tenido algún problema.

2) Compilación e integración de datos

Una vez extraída la información MIAPE, la herramienta permite el análisis e inspección de los datos de identificación de péptidos y proteínas de cada uno de los documentos MIAPE MSI. Sin embargo, no olvidemos que cada uno de estos documentos proviene de los resultados de una única búsqueda en la base de datos. Muchas veces los experimentos de caracterización de proteomas están precedidos de una separación cromatográfica previa, por lo que las fracciones resultantes se analizan independientemente en el espectrómetro y posteriormente en los motores de búsqueda. Por tanto al final, tendremos tantos documentos MIAPE MS y MSI como fracciones en el experimento. Es por tanto necesaria la integración de todos esos datos. Para ello, la herramienta mostrará al usuario todos los documentos MIAPE MSI disponibles en su ámbito y podrá seleccionar los que sean de su interés para compararlos, inspeccionarlos o agregarlos a otros experimentos. En la sección 3.2.3, ya explicamos el funcionamiento de la compilación o agregación de datos que permite la herramienta MIAPE Extractor, creándose un proyecto de inspección en el que se organizan los documentos MIAPE MSI en diferentes niveles de un árbol (Figura 10).

Además, y como también se mencionó en la sección 3.2.3, se implementó el método de agrupamiento de proteínas PAnalyzer (Prieto, Aloria et al. 2012), aplicándose a todos los datos dependientes de un nodo definido en el proyecto. Esto permite el control y etiquetado de las diferentes ambigüedades inherentes a la inferencia de proteínas a partir de la identificación de péptidos, lo cual nos permite saber, por ejemplo, si realmente hemos detectado una isoforma de una familia proteica o si por el contrario, con los péptidos identificados no somos capaces de distinguir entre los miembros de dicha familia. En el ámbito de la iniciativa española del Proyecto del Proteoma Humano (HPP) esta incorporación ha sido crucial para el estudio de las

diferentes proteo-formas derivadas de los genes del cromosoma 16 (Segura, Medina-Aunon et al. 2013, Segura, Medina-Aunon et al. 2013).

Recientemente se ha modificado la herramienta para poder introducir sets de datos externos, por ejemplo, publicados en una revista científica, en la que no tenemos la posibilidad de obtener un fichero mzIdentML estándar, sino que tenemos una tabla Excel con las identificaciones. Esta extensión de la herramienta lee un fichero de texto, de fácil creación a partir de cualquier tabla en la que una columna contiene los números de acceso de las proteínas, otra columna tiene las secuencias peptídicas y otra columna contiene los valores de puntuación de cada uno de los PSMs. La estructura de este fichero es analizada para construir un fichero MIAPE MSI XML con dicha información, pudiéndose utilizar en el posterior análisis de la misma manera que si fuese un documento MIAPE MSI completo.

3) Filtrado de datos

Una vez definido el proyecto de inspección, los datos pueden ser cribados a partir de un conjunto de filtros de diferentes tipos:

- Filtro por FDR o tasa de error: en el caso de que las búsquedas hayan sido identificadas con una base de datos señuelo, este filtro permite conservar un conjunto de datos con una tasa de error seleccionada, tanto a nivel de PSM, de péptido o a nivel de proteína. Los detalles del cálculo de la tasa de error se explicaron en la sección 3.1.4.
- Filtro por valor de corte de puntuación: este filtro permite descartar las identificaciones con un valor de puntuación mayor o menor que un valor introducido.
- Filtro por longitud de péptido: permite descartar los péptidos que tengan una longitud en su cadena de aminoácidos menor a un valor dado. Este filtro se aplica también por defecto a péptidos con menos de 7 aminoácidos (aunque puede modificarse) para evitar el caso de que una misma secuencia peptídica pertenezca a una proteína normal y otra señuelo, lo cual podría desvirtuar la tasa de error del conjunto de datos.
- Filtro por modificación de péptidos: permite conservar únicamente los péptidos y proteínas que contengan una modificación post-traduccional indicada.
- Filtro por lista de identificadores de proteínas: este filtro permite centrar el análisis únicamente en los datos relativos a unas proteínas de interés conocidas.
- Filtro por lista de secuencias peptídicas: análogo al anterior, filtra los datos para mostrar únicamente la información relativa a una lista de péptidos introducidos por el usuario.
- Filtro por ocurrencia: permite descartar péptidos o proteínas que no se hayan detectado un número mínimo de veces.

- Filtro por número de péptidos asignados a cada proteína: permite descartar las proteínas que no hayan sido identificadas con un mínimo número de péptidos.
- Filtro para seleccionar péptidos candidatos para un análisis dirigido MRM (*Multiple Reaction Monitoring*): permite seleccionar los péptidos con ciertas características de secuencia y longitud, los cuales son aptos para un análisis cuantitativo por MRM.

4) Inspección de datos

El módulo de inspección de la herramienta MIAPE Extractor fue desarrollado para poder aprovechar la riqueza y gran cantidad de datos contenidos en los documentos MIAPE completos. Así pues, por medio de una variedad enorme de gráficas, los datos definidos en el proyecto de inspección son visualizados de manera fácil e intuitiva por cualquier usuario. Además, ya que en el proyecto se pueden seleccionar datos provenientes de diferentes experimentos, e incluso datos provenientes de tablas de identificaciones publicadas, la herramienta MIAPE Extractor constituye una de las herramientas más útiles para comparar datos provenientes de diferentes aproximaciones o laboratorios, siendo la única capaz de integrarlos.

Algunos de los aspectos que se pueden monitorizar son: número de identificaciones, número de identificaciones exclusivas a cada conjunto de datos comparado, solapamiento de identificaciones, sensibilidad y precisión, curvas de FDR, distribución de puntuaciones, comparación de puntuaciones, distribución de cortes enzimáticos omitidos (*missed cleavages*), distribución de longitudes de péptidos, número de péptidos por proteína, distribución de modificaciones post-traduccionales, distribución de palabras de las descripciones de las proteínas, heat maps de identificaciones, distribución de puntos isoeléctricos, etc...

En la siguiente sección se mostrarán los datos inspeccionados para el experimento multi-centro 6 de ProteoRed (PME6).

5) Envío de datos a ProteomeXchange

Una vez filtrados e inspeccionados los datos, éstos pueden ser exportados fácilmente en tablas Excel, o bien se pueden exportar los datos en formato PRIDE XML. Los ficheros PRIDE XML generados, contendrán toda la información MIAPE siempre y cuando los mzIdentML estuviesen completos (en todos los casos conocidos lo estaban) y los MIAPE MS fuesen creados utilizando una plantilla de metadatos completa. Así pues, estamos hablando de ficheros PRIDE XML que cumplen las directrices MIAPE, siendo la única herramienta capaz de anotar tan detalladamente dichos ficheros.

Los ficheros PRIDE XML pueden ser enviados directamente al repositorio público EBI-EMBL (*European Bioinformatics Institute – European Molecular Biology Laboratory*) PRIDE. Sin embargo, la herramienta proporciona la manera de preparar un envío más completo aún, esto es, un envío completo a ProteomeXchange. Así pues, simplemente pinchando en el botón "ProteomeXchange" de la interfaz de inspección de datos, se muestra una pantalla para introducir los metadatos necesarios para el envío (título y descripción del experimento o de los datos y una lista de palabras clave relacionadas) (Figura 51).

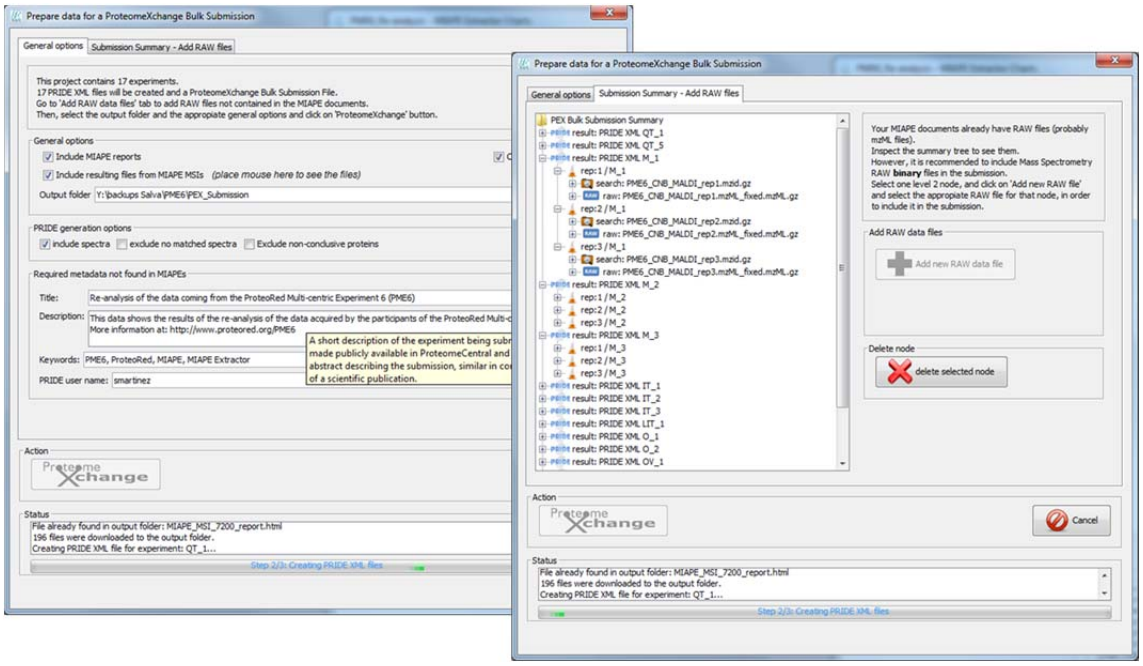


Figura 51. Interfaz gráfica del MIAPE Extractor para la preparación del envío de los datos a ProteomeXchange: A la izquierda, existen una serie de opciones como si incluir o no los informes MIAPE legibles o los ficheros mzIdentML, incluir o no los espectros en los ficheros PRIDE, etc… Además, tres formularios permiten introducir el título y la descripción del proyecto, así como una lista de palabras clave. A la derecha, en la pestaña de "*Submission Summary*", se muestra un árbol con todos los ficheros que se van a incluir en el envío de los datos. Adicionalmente, se pueden añadir manualmente ficheros de datos crudos (RAW) asociándolos a cada uno de los nodos de nivel 2 del árbol.

Después, tras pulsar el botón "ProteomeXchange", el proceso terminará tras el término de tres tareas:

- los ficheros necesarios se empezarán a bajar en el directorio local de destino, incluyendo también, en este caso, los ficheros de datos crudos referenciados en los MIAPE. Si se selecciona la opción de comprimir los ficheros en formato gzip, se comprimirán. Estos ficheros son:
    - Ficheros PRIDE XML (uno por cada nodo de nivel 1).
    - Ficheros de datos crudos. Estos ficheros no se recogen de por sí en el flujo de trabajo proporcionado por la herramienta MIAPE Extractor. Así que la

herramienta en este punto permite introducirlos manualmente asociándolos a cada una de los MIAPE MSI (nodos de nivel 2).

- Ficheros mzIdentML (utilizados para crear cada documento MIAPE MSI).
- Ficheros mgf o mzML (utilizados para crear cada documento MIAPE MS).
- Informes de documentos MIAPE MS y MSI (ficheros HTML con los documentos MIAPE legibles).

- Se crearán los ficheros PRIDE XML de cada experimento (de cada nodo de nivel 1) siguiendo las opciones seleccionadas por el usuario.
- Se creará el fichero resumen del envío a ProteomeXchange, con los metadatos necesarios, y relacionando todos los ficheros entre sí (Figura 52), que contendrá además los metadatos requeridos extraídos de los documentos MIAPE, como los espectrómetros, las modificaciones post-traduccionales, la especie a la que pertenece la muestra, etc.

```
MTD  name Alberto Medina-Aunion
MTD  email jamedina@cnb.csic.es
MTD  affiliation    Centro Nacional de BiotecnologÃ-a, CSIC, Madrid, Spain.
MTD  title  Chr16-HPP. Shotgun Analysis improvement. JPR HPP Special issue
MTD  description  The Chromosome 16 Consortium is integrated in the Human Proteome Project that aims to develop an entire map of the proteins encoded by the human genome following a
MTD  keywords   HPP,MIAPE,Networking,Chromosome centric
MTD  type  SUPPORTED
MTD  species    [NEWT,9606,Homo Sapiens (Human),]
MTD  instrument  [PSI-MS,MS:1001541,maXis,]
MTD  instrument  [PSI-MS,MS:1001742,LTQ Orbitrap Velos,]
MTD  instrument  [PSI-MS,MS:1000639,LTQ Orbitrap XL ETD,]
MTD  instrument  [PSI-MS,MS:1000449,LTQ Orbitrap,]
MTD  instrument  [PSI-MS,MS:1000932,TripleTOF 5600,]
MTD  modification       [PSI-MOD,MOD:00052,Acetyl,]
MTD  modification       [PSI-MOD,MOD:01060,carbamidomethylc,]
MTD  modification       [PSI-MOD,MOD:00040,Gln->pyro-Glu,]
MTD  modification       [PSI-MOD,MOD:00412,oxidation,]
MTD  pride_login  jamedina

FMH  file_id    file_type    file_path    file_mapping
FME  1    result X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\SPHPP_VHIO_MCF7_QTOF_GEL_curated_FDR1_protein.xml    2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,2
FME  2    raw  X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\RawData_JPR_SpecialIssue_2013\VHIO\MCF7_Gel\SPHPP_VHIO_MCF7_QTOF_GEL_R1_1_15.baf    3,4,5,6
FME  3    peak  X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\SPHPP_VHIO_MCF7_QTOF_GEL_R1_1_15.mgf.gz  4,5,6
FME  4    other X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\MIAPE_MS_6714_report.html
FME  5    search       X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\SPHPP_VHIO_MCF7_QTOF_GEL_R1_1_15.mzid.gz  6
FME  6    other X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\MIAPE_MSI_6609_report.html
FME  7    raw  X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\RawData_JPR_SpecialIssue_2013\VHIO\MCF7_Gel\SPHPP_VHIO_MCF7_QTOF_GEL_R1_3_15.baf    8,9,10,11
FME  8    peak  X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\SPHPP_VHIO_MCF7_QTOF_GEL_R1_3_15.mgf.gz  9,10,11
FME  9    other X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\MIAPE_MS_6717_report.html
FME  10   search       X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\SPHPP_VHIO_MCF7_QTOF_GEL_R1_3_15.mzid.gz  11
FME  11   other X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\MIAPE_MSI_6610_report.html
FME  12   raw  X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\RawData_JPR_SpecialIssue_2013\VHIO\MCF7_Gel\SPHPP_VHIO_MCF7_QTOF_GEL_R1_4_15.baf    13,14,15,16
FME  13   peak  X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\SPHPP_VHIO_MCF7_QTOF_GEL_R1_4_15.mgf.gz  14,15,16
FME  14   other X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\MIAPE_MS_6718_report.html
FME  15   search       X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\SPHPP_VHIO_MCF7_QTOF_GEL_R1_4_15.mzid.gz  16
FME  16   other X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\MIAPE_MSI_6611_report.html
FME  17   raw  X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\RawData_JPR_SpecialIssue_2013\VHIO\MCF7_Gel\SPHPP_VHIO_MCF7_QTOF_GEL_R1_5_15.baf    18,19,20,21
FME  18   peak  X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\SPHPP_VHIO_MCF7_QTOF_GEL_R1_5_15.mgf.gz  19,20,21
FME  19   other X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\MIAPE_MS_6719_report.html
FME  20   search       X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\SPHPP_VHIO_MCF7_QTOF_GEL_R1_5_15.mzid.gz  21
FME  21   other X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\MIAPE_MSI_6612_report.html
FME  22   raw  X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\RawData_JPR_SpecialIssue_2013\VHIO\MCF7_Gel\SPHPP_VHIO_MCF7_QTOF_GEL_R1_6_15.baf    23,24,25,26
FME  23   peak  X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\SPHPP_VHIO_MCF7_QTOF_GEL_R1_6_15.mgf.gz  24,25,26
FME  24   other X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\MIAPE_MS_6720_report.html
FME  25   search       X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\SPHPP_VHIO_MCF7_QTOF_GEL_R1_6_15.mzid.gz  26
FME  26   other X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\MIAPE_MSI_6613_report.html
FME  27   raw  X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\RawData_JPR_SpecialIssue_2013\VHIO\MCF7_Gel\SPHPP_VHIO_MCF7_QTOF_GEL_R1_7_15.baf    28,29,30,31
FME  28   peak  X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\SPHPP_VHIO_MCF7_QTOF_GEL_R1_7_15.mgf.gz  29,30,31
FME  29   other X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\MIAPE_MS_6721_report.html
FME  30   search       X:\ftpEstrellaPolar\proteomica\PX_Submission_Jul13\MCF7\SPHPP_VHIO_MCF7_QTOF_GEL_R1_7_15.mzid.gz  31
```

Figura 52. Fichero resumen del envío de datos a ProteomeXchange. El fichero de texto contiene una serie de líneas de cabecera con ciertos metadatos como los datos de contacto, palabras clave, título y descripción del proyecto, especie a la que pertenece la muestra, espectrómetros utilizados, etc. Más abajo, todos los ficheros generados por la herramienta MIAPE Extractor se clasifican por tipos de ficheros (*result, raw, peak, search u* other) y se relacionan entre sí por medio de referencias entre ellos.

Dicho fichero podrá ser utilizado directamente en la herramienta de envío de datos del ProteomeXchange (como una *bulk submission*) y el usuario únicamente tendrá que dar a un botón para enviar todos sus experimentos de una sola vez (Figura 53). Así pues, es posible el

envío de datos a ProteomeXchange siguiendo las directrices MIAPE de forma fácil e intuitiva, lo cual es únicamente posible, hasta el momento, gracias a nuestro flujo de trabajo.

Figura 53. Interfaz gráfica de la herramienta de envío de datos al ProteomeXchange: En rojo se resalta el botón que hay que pinchar para iniciar una "*bulk submission*", para la cual es necesario el fichero resumen del envío que proporciona la herramienta MIAPE Extractor.

Para demostrar la utilidad y potencialidad de la herramienta, en el Anexo se describe el reanálisis de los datos de espectrometría de masas adquiridos por los participantes del estudio multicentro 6 de ProteoRed. Tanto los datos enviados (Anexo A) como los datos reanalizados (Anexo B) se procesaron en la herramienta MIAPE Extractor. El reanálisis de los datos, usando un mismo motor de búsqueda, Mascot, y una misma base de datos, permite evaluar de forma más robusta el rendimiento de cada una de las plataformas, eliminándose posibles varianzas debidas a diferencias en los flujos de trabajo y análisis.

4.2.6. Definición de las directrices MIAPE para experimentos cuantitativos en Proteómica, dentro del marco de trabajo del HUPO-PSI.

A lo largo de esta tesis se ha hablado principalmente de los datos de identificación de péptidos y proteínas provenientes de experimentos de identificación a gran escala. Las mejoras metodológicas y tecnológicas aparecidas recientemente en el campo de la espectrometría de masas, han hecho posible que la cuantificación basada en espectrometría de masas emerja como una de las técnicas más utilizadas para cuantificar proteínas, dando lugar a estudios cuantitativos a gran escala. Existen numerosas estrategias para la cuantificación absoluta o relativa de las proteínas, así como para el análisis estadístico asociado necesario para discernir las proteínas significativamente diferenciales en un estudio comparativo, cada una de los cuales utilizan diferentes métricas y por tanto, diferentes flujos de trabajo de análisis. Dada esta gran heterogeneidad de nuevas aproximaciones Proteómicas emergentes en los últimos años y dado que los estándares relacionados con toda la parte de espectrometría de masas y el análisis de datos para la identificación de péptidos y proteínas estaban lo suficientemente asentados dentro de la comunidad científica proteómica (Binz, Barkovich et al. 2008, Deutsch 2008, Taylor, Binz et al. 2008, Eisenacher 2011, Martens, Chambers et al. 2011, Jones, Eisenacher et al. 2012), se hizo evidente la necesidad del desarrollo de estándares de datos y directrices MIAPE para la Proteómica cuantitativa. El estándar mzQuantML ha sido recientemente publicado (Walzer, Qi et al. 2013) como el estándar para la representación de diferentes tipos de datos proteómicos cuantitativos. Paralelamente, y siendo parte del trabajo de esta tesis, se dirigió una iniciativa apoyada por el HUPO-PSI para la definición de las directrices MIAPE para la Proteómica cuantitativa. Para ello, se coordinó un grupo de expertos con amplia experiencia en diferentes técnicas cuantitativas, pertenecientes todos a la red de laboratorios de Proteómica de ProteoRed, para la elaboración primeramente de un borrador de lo que sería el nuevo módulo MIAPE Quant. Ese borrador fue el resultado de la fusión a su vez de varios módulos específicos elaborados para describir diferentes técnicas cuantitativas: aproximaciones basadas en marcaje como ICPL (*Isotope-coded protein label*), iTRAQ (*Isobaric tag for relative and absolute quantification*) o SILAC (*Stable isotope labelling by amino acids in cell culture*), aproximaciones sin marcaje como el contaje de espectros (*Spectral counting*) o la cuantificación basada en la intensidad del ion precursor, o técnicas de cuantificación dirigida como MRM (*Multiple reaction monitoring*). Dicho borrados fue revisado a su vez por los integrantes de la reunión anual del HUPO-PSI en Heidelberg en 2011. Luego, tras diversas revisiones y tras elaborar el documento final junto con diez documentos de ejemplo describiendo experimentos cuantitativos reales, el nuevo módulo se envió al proceso de revisión formal del HUPO-PSI, aceptándose como documento "final" en Octubre de 2012 (http://www.psidev.info/miape-quant-

1.0). Posteriormente fue revisado, aceptado y publicado en un número especial de la revista *Journal of Proteomics Research* llamado "Estandarización y control de calidad" (*Standardization and quality control*) (Martinez-Bartolome, Deutsch et al. 2013) (Figura 54).

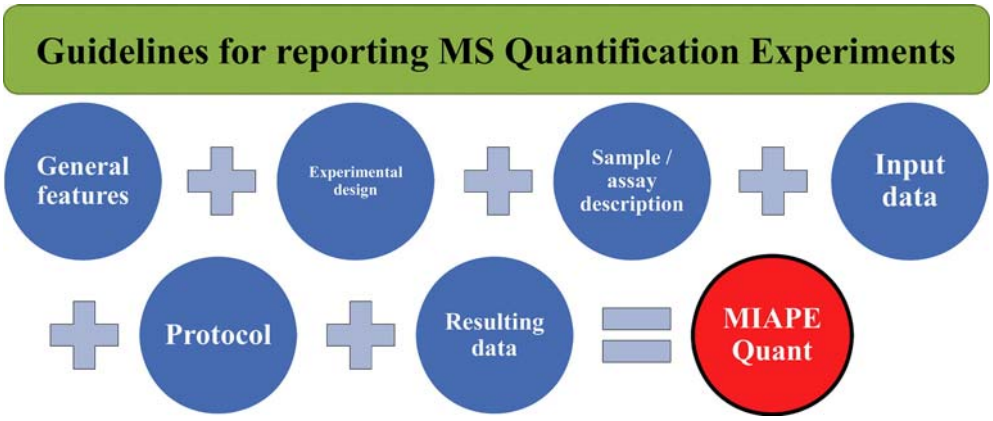


Figura 54. Esquema de las directrices MIAPE sobre experimentos proteómicos cuantitativos basados en espectrometría de masas.

Discusión

5. Discusión

5.1. Desarrollo de los métodos para la validación de identificaciones de péptidos a gran escala

Casi todos los métodos descritos para hacer inferencias estadísticas sobre la identificación de péptidos usando el motor de búsqueda SEQUEST se basan en el análisis del conjunto de las mejores puntuaciones (XCorr) obtenidas de buscar una colección de espectros contra una base de datos de secuencias de proteínas (Keller, Nesvizhskii et al. 2002, MacCoss, Wu et al. 2002, Moore, Young et al. 2002, Kislinger, Rahman et al. 2003, Lopez-Ferrer, Martinez-Bartolome et al. 2004). Las mejores puntuaciones se agrupan y se ordenan para construir una distribución acumulativa de puntuaciones, la cual se usa para determinar la significación estadística de las identificaciones y también para calcular la FDR, si se repite la búsqueda con una base de datos señuelo. Estas distribuciones acumulativas se pueden concebir como una estimación estadística de la distribución promedio de probabilidad de la mejor puntuación (*average score distribution*), siendo una función que evalúa la probabilidad de obtener un espectro de entre toda la colección de espectros al que se le asigna una puntuación igual o mejor que la puntuación observada.

Otra aproximación válida para la evaluación de las identificaciones se basa en el análisis de la significatividad individual de cada una de las asignaciones entre espectros y secuencias, evaluando la probabilidad de que la puntuación obtenida por un espectro sea igual o mejor que la que se obtendría al buscar ese espectro contra una colección de secuencias candidatas. Esta probabilidad se calcula por medio de las distribuciones individuales (*single-spectrum distributions*) de puntuaciones.

En el trabajo aquí expuesto y publicado en el 2008 (Martinez-Bartolome, Navarro et al. 2008), realizamos un estudio analítico sobre las características generales de las diferentes distribuciones de probabilidad obtenidas al buscar colecciones de espectros contra bases de datos de secuencias de proteínas. Para ello analizamos las distribuciones individuales de espectros y expresamos la distribución promedio como la combinación de las distribuciones individuales (esta parte del modelo teórico fue realizada por el Dr. Fernando Martín-Maroto), debido principalmente a que las distribuciones individuales pueden ser muy diferentes para diferentes espectros, por lo que no se puede esperar que grandes conjuntos de datos tengan un comportamiento estadístico homogéneo.

Una de las demostraciones realizadas en este trabajo de los conceptos teóricos del modelo es que el valor de la distribución promedio para una cierta puntuación indica la fracción de espectros del conjunto total que tienen una calidad mejor, es decir, que tienden a producir una puntuación mejor por azar. Este término de "calidad" ha sido aplicado previamente a espectros MS/MS refiriéndose a ello como una propiedad de los espectros que indica la probabilidad de que efectivamente se generen por la fragmentación de péptidos reales (Bern, Goldberg et al. 2004, Strittmatter, Kangas et al. 2004). En este estudio, sin embargo, el término calidad es entendido como una propiedad de los espectros que hace que obtengan una puntuación mayor al enfrentarlos contra bases de datos señuelo.

Así pues, y de acuerdo con nuestro análisis, se espera que la forma de las distribuciones promedio de probabilidad refleje la distribución de calidades de la población de espectros; por tanto dependerán principalmente de los factores que alteren la calidad de las fragmentaciones. Por ejemplo, aumentando la concentración de péptidos en la muestra, se incrementará la proporción de espectros con información de secuencia, por lo que la distribución de calidades cambiará, ya que esos espectros tenderán a producir por azar mejores puntuaciones. Por otro lado, cambiar el tamaño de la base de datos cambiando el número de secuencias candidatas que se comparan con cada espectro, afectará también a la distribución de puntuaciones, siendo particularmente evidente en la cola correspondiente a la región de baja probabilidad, como describen las ecuaciones de escalado.

Por tanto, nuestro análisis demuestra que es imposible establecer un criterio universal de corte para determinar la significación de las asignaciones de los péptidos basándose en la mejor y la segunda mejor puntuación; ya que estos criterios deben establecerse dependiendo de cada experimento. De la misma forma, no se pueden hacer modelos estadísticos empíricos a partir de distribuciones predefinidas o universalmente establecidas, ya que se espera que cambien, reflejando así la distribución de calidades asociada a cada caso en particular. El análisis que presentamos en este trabajo proporciona por primera vez unos conceptos básicos para describir y predecir estos fenómenos.

Otra consecuencia de gran utilidad práctica derivada de nuestro trabajo es que la cola de las distribuciones promedio de probabilidad se puede predecir gracias a la ecuación de escalado, así que es posible estimar cuál sería la probabilidad asociada a una asignación de un péptido si la búsqueda se hiciese con una base de datos diferente.

De acuerdo a nuestros resultados, la razón de probabilidad constituye uno de los métodos más atractivos cuando las distribuciones de probabilidad de grandes conjuntos de espectros se analizan de manera conjunta, considerando únicamente la información dada por la mejor y la segunda mejor puntuación. Esta razón de probabilidad debe ser entendida como una

probabilidad condicionada, que tiene en cuenta la información de calidad que proporciona la segunda mejor puntuación y por tanto, como se ha demostrado a lo largo de este trabajo, corrige la desviación intrínseca derivada de la calidad de los espectros. Además, destaca por la sencillez de su aplicación en la práctica, tanto conceptual como computacionalmente, evitándose el uso de parámetros de ajuste a una distribución o extrapolaciones en la cola de la distribución promedio de las puntuaciones aleatorias. No se introduce por tanto ningún tipo de imprecisión en el cálculo de las probabilidades promedio de cada puntuación, y los valores de FDR se calculan de manera directa para cada uno de los PSMs. Además, tiene otra gran ventaja con respecto a otros métodos, como es el hecho de no ser necesario separar y analizar la población de espectros de acuerdo a parámetros como su longitud o su carga, al igual que no es necesario el uso de funciones matemáticas predefinidas ni, como hemos dicho, encontrar parámetros de ajuste óptimos a las distribuciones experimentales. A pesar de su simplicidad, el método de la razón de probabilidad proporciona un rendimiento superior o al menos comparable al de los métodos estadísticos empíricos publicados anteriormente.

Aunque por lo general se espera que la mejor segunda puntuación sea consecuencia de una asignación aleatoria, este no es el caso de ciertas situaciones en las que se puede esperar que una gran proporción de las segundas mejores puntuaciones se deriven de asignaciones a secuencias con cierta homología con respecto a las secuencias peptídicas reales. Esta situación ocurre por ejemplo, cuando se usan bases de datos sin filtros por taxonomía, o que incluyan diferentes isoformas de una misma proteína. En estos casos cobra gran utilidad el hecho de que la fracción de calidad se puede calcular también a partir de las distribuciones promedio de la tercera, cuarta, etc… mejores puntuaciones. Por tanto es posible calcular la razón de probabilidad usando por ejemplo, la mejor y la cuarta mejor puntuación (en vez de la segunda mejor), o incluso utilizar una calidad promedio calculada a partir de la tercera, la cuarta y la quinta mejor puntuación.

El método de calidad única por su parte debería verse como un ejemplo simple y directo de cómo aplicar en la práctica el concepto de la calidad de un espectro. Como ya se ha expuesto anteriormente, las ventajas de éste método no sólo aparecen cuando se compara con los métodos estadísticos empíricos, sino que también permite analizar cada uno de los espectros de manera independiente e individualizada; no es necesario construir las distribuciones estadísticas promedio de las puntuaciones usando bases de datos aleatorias; permite el cálculo directo de la FDR y las distribuciones predichas cumplen perfectamente las ecuaciones de escalado, por lo que son totalmente predecibles. Aunque con este método se aumenta la complejidad del análisis y, por tanto, el tiempo de computación, consideramos que es un método idóneo cuando se utilizan bases de datos pequeñas, debido a su robustez conceptual y predictibilidad. De hecho, creemos que el concepto probabilístico que subyace a este método puede constituir una base para nuevos motores de búsqueda universales y no empíricos.

Aunque la validez de nuestro análisis matemático haya sido probada estudiando el comportamiento de las distribuciones de las puntuaciones usando uno de los motores de búsqueda más populares, SEQUEST, es importante señalar que el análisis matemático y las consecuencias que se derivan de él son, en general, aplicables a cualquier esquema de puntuación que pueda tener cualquier otro motor de búsqueda. De hecho este método ha sido aplicado también a los resultados del motor de búsqueda Mascot obteniéndose unos resultados razonables (no mostrados). Pensamos, por tanto, que la aproximación matemática desarrollada aquí es más valiosa por sí misma que los dos métodos derivados de los análisis de la teoría de los resultados de SEQUEST, los cuales pueden verse más bien como ejemplos para enseñar la utilidad práctica de este estudio. Pensamos que nuestro trabajo abre nuevos caminos para el desarrollo de nuevos algoritmos de búsqueda susceptibles de producir criterios normalizados y universales para la identificación de péptidos.

Finalmente, el presente estudio proporciona varias posibilidades que podrían ayudar a definir un criterio general para validar las asignaciones de péptidos. Tanto el método de la razón de probabilidad como el método de la calidad única, cumpliendo la ecuación de escalado, permiten calcular una probabilidad promedio normalizada, lo cual permitiría comparar los resultados obtenidos con diferentes condiciones o por diferentes laboratorios.

5.2. Desarrollo de herramientas basadas en estándares

Los estándares de representación de datos definidos por el HUPO-PSI han permitido a la comunidad científica la posibilidad de la reutilización de datos externos, así como la comparación de información proveniente de diferentes fuentes. En definitiva, han proporcionado un lenguaje común para así compartir datos e información, dando respuesta al gran problema existente con la gran heterogeneidad y la creciente y enorme cantidad de datos proteómicos que se generan año tras año, que de otra forma resultan difícilmente comparables.

Por su parte, las directrices MIAPE definidas también por el HUPO-PSI (Taylor, Paton et al. 2007) proporcionan un estándar de calidad que asegura la clara y precisa interpretación de los resultados de los experimentos proteómicos, así como la información necesaria para poder reproducirlos en la medida de lo posible. Sin embargo, pese a su clara utilidad como valor añadido de calidad, es una cuestión de que las revistas científicas, o también, de que los revisores de dichas revistas, adopten las directrices MIAPE como la información mínima que debe incluir una publicación de un experimento proteómico. Sin embargo, pese a que algunas revistas especializadas en Proteómica como *Proteomics* o *Molecular and Cellular Proteomics* recomiendan en sus instrucciones para autores el seguimiento de las directrices MIAPE, no ha llegado aún el momento en el que las revistas adopten las directrices MIAPE literalmente como la información mínima a incluir en sus publicaciones, unificando así unos criterios de mínimos para toda la comunidad científica. Una de las razones aportadas por algunos de los editores de las revistas es que no existen las herramientas necesarias que permitan la comprobación automática del seguimiento de estas directrices en los resultados enviados a publicar. Otra de las razones puede ser el difícil compromiso entre practicidad y suficiencia en la que se deben basar las directrices MIAPE, siendo un equilibrio que claramente, en la opinión de las revistas, se debe inclinar más hacia la practicidad. Su argumentación, totalmente comprensible, es que las directrices MIAPE no deben entorpecer el proceso de publicación de trabajos científicos: ni para los autores que deben incluir toda la información en el manuscrito, ni para los revisores o el editor, que deben de comprobar si alguna información requerida no está presente. Es por ello por lo que los editores de las revistas reclaman herramientas que ayuden a mitigar el sobreesfuerzo que supone para ambas partes la incorporación de las directrices MIAPE como propias. Gran parte del trabajo mostrado en esta tesis se ha enfocado en esa dirección y ha conseguido que ahora numerosos revisores exijan la adecuación a las directrices MIAPE respaldándose en nuestras herramientas. Sin embargo, nos queda bastante trabajo para asegurar un rendimiento óptimo de nuestros sistemas a escala mundial, lo que nuestros recursos actuales nos han impedido alcanzar, para así conseguir que oficialmente las revistas científicas adopten unas directrices comunes en el campo de la Proteómica.

En este trabajo se ha demostrado la utilidad de la utilización de los dos tipos de estándares desarrollados por HUPO-PSI, los de representación de datos y las directrices MIAPE, desarrollándose varios tipos de herramientas bioinformáticas basadas en ellos.

La herramienta online generadora de MIAPEs, desarrollada sobre el repositorio de experimentos proteómicos de ProteoRed, constituye el único recurso existente para la creación de documentos MIAPE descriptores de experimentos, así como el único repositorio de este tipo de acceso público. El único recurso que existió muy similar a éste, el *MIAPEGelDB* aunque únicamente dirigido a la creación de documentos MIAPE GE (*Gel Electrophoresis*), ya no está soportado por sus desarrolladores y únicamente se pueden consultar los documentos allí almacenados (Robin, Hoogland et al. 2008). Nuestra herramienta puede ayudar al proceso de revisión de un artículo: los autores pueden publicar sus experimentos siguiendo las directrices MIAPE y pueden adjuntar los informes en el envío del manuscrito a la revista, como parte de la sección de materiales y métodos o como material suplementario.

Posteriormente, y gracias a la aparición de los estándares de representación de datos mzML y mzIdentML para los datos de espectrometría de masas y de identificaciones respectivamente, la automatización del proceso de descripción y publicación de experimentos siguiendo las directrices MIAPE fue por fin posible. Esta automatización se hizo realidad gracias, primero, al desarrollo de la librería de código abierto ProteoRed Java MIAPE API, que permite de forma programática extraer la información MIAPE de los ficheros de datos proteómicos estándares, así como exportar la información MIAPE en formato PRIDE XML, y en segundo lugar a la aparición de nuevas librerías en Java para el manejo de estos y otros ficheros de datos proteómicos como jmzIdentML, jmzML, XTandem parser, PRIDE core library, etc. Es por tanto un esfuerzo de la comunidad bioinformática Proteómica gracias al cual, hoy en día, el desarrollo de nuevos métodos y herramientas es mucho más fácil que hace tan sólo 5 años.

Aprovechando todos estos trabajos, (usando la librería de manejo de información MIAPE, librerías de manejo de ficheros proteómicos, y el acceso remoto y seguro de los servicios web desplegados), se desarrolló finalmente una herramienta, el ProteoRed MIAPE Extractor, que ha ido creciendo paulatinamente hasta convertirse en una herramienta de manejo, integración, análisis y envío de datos, proporcionando un flujo completo de análisis bioinformático basado en estándares y en las directrices MIAPE.

Los análisis e inspección de los datos de identificación de péptidos y proteínas por parte del MIAPE Extractor son sin duda uno de los más potentes, si no el más potente, de las herramientas existentes, en el aspecto de la gran cantidad de datos que es capaz de integrar, analizar y filtrar. La integración de los resultados de experimentos con pre-fraccionamiento en los cuales tenemos una búsqueda en bases de datos por cada banda de gel o fracción

cromatográfica, es totalmente automática, incluso reorganizando las asignaciones péptido-proteína por medio del algoritmo de agrupamiento de proteínas PAnalyzer (Prieto, Aloria et al. 2012). Además, la herramienta proporciona una gran flexibilidad y versatilidad para agrupar los datos organizándolos de manera jerárquica, lo que permite su integración e inspección a distintos niveles en un mismo tiempo. Asimismo, y para cada uno de los niveles o nodos en la jerarquía de datos definida por el usuario, existen numerosísimas gráficas que permiten la inspección completa de los datos de una forma fácil y rápida (ver Anexo).

Además de poder inspeccionar los datos, la herramienta proporciona numerosos posibilidades para filtrar los datos, ya sea por puntuaciones o por tasas de error o para centrarse en el análisis de un subconjunto de los datos, como péptidos con una modificación post-traduccional concreta o por ejemplo, proteínas que se han visto más de una vez a lo largo de las réplicas inspeccionadas.

Es de gran relevancia el paso final dentro del flujo de trabajo, que consiste en la preparación de los datos para un envío a ProteomeXchange. Como hemos comentado, ProteomeXchange proporciona una entrada común para el envío de datos proteómicos y un portal público para acceder a ellos. Existe una herramienta, desarrollada por varios grupos del consorcio del ProteomeXchange, que permite de forma intuitiva incluir los ficheros necesarios y los metadatos mínimos para realizar un envío de datos correcto a ProteomeXchange. Sin embargo, cuando el envío se compone de varios experimentos, por ejemplo, con una separación previa a la adquisición, el número de ficheros a crear, recopilar, y a interrelacionar entre sí, para que el envío sea coherente requiere un trabajo previo manual muy grande. Nuestra herramienta es capaz de recopilar todos los ficheros necesarios para el envío, además de crear los ficheros PRIDE XML con todos los metadatos incluidos en los documentos MIAPE e incluye adicionalmente los ficheros de resultados originales de las búsquedas en mzIdentML o los informes MIAPE en formato HTML tanto MS como MSI.

Debido a su gran utilidad y robustez, la herramienta ProteoRed MIAPE Extractor es la herramienta idónea para el análisis, comparación o integración de grandes cantidades de datos. Por ello, está siendo utilizada por el consorcio español del Proyecto del Proteoma Humano (Sp-HPP) para recopilar e integrar los resultados de identificación de péptidos y proteínas en diversas líneas celulares provenientes de diferentes laboratorios españoles usando diferentes plataformas de espectrometría de masas. Cada uno de esos laboratorios utiliza la herramienta para extraer la información MIAPE de sus experimentos y almacenarla en nuestro repositorio. Una vez allí, es recopilada y analizada con el MIAPE Extractor de manera centralizada, integrando todos los resultados a la vez y extrayendo la información de interés para el proyecto (Segura, Medina-Aunon et al. 2013, Segura, Medina-Aunon et al. 2013).

Como ejemplo de uso de la herramienta ProteoRed MIAPE Extractor se muestra en el anexo, para el análisis de los datos provenientes de unos de los experimentos multi-centro organizados por ProteoRed, el PME-6. En el Anexo se puede ver el flujo completo de integración, análisis, filtrado, inspección y envío a ProteomeXchange del reprocesado de los datos de identificación de 17 laboratorios con 3 réplicas cada uno de ellos, así como también la integración y comparación de los propios datos enviados por cada participante en las plantillas de resultados. Con ello se quiere demostrar la utilidad de la herramienta y su facilidad de uso para el manejo y análisis de grandes cantidades de datos, tareas que manualmente, o usando herramientas como tablas Excel sería prácticamente imposible. El envío de los datos reprocesados al repositorio público ProteomeXchange demuestra también que el MIAPE Extractor constituye hoy en día un recurso adicional para compartir los datos a través de ProteomeXchange, siendo realmente la única herramienta capaz de hacerlo asegurando el cumplimiento de las directrices MIAPE.

Por otro lado, las herramientas desarrolladas para la validación semántica y MIAPE de los ficheros de representación de datos estándares mzML y mzIdentML son de gran utilidad para el desarrollo de nuevos software para el manejo y creación de dichos ficheros, proporcionando una herramienta para validarlos y la información contenida en los informes de validación será de gran utilidad para poder mejorar o ampliar la información anotada en ellos. Además, pese a que nuevos vocabularios controlados se incluyan en la ontología, los validadores se mantendrán siempre actualizados y los tendrán siempre en cuenta debido a que utilizan el servicio remoto de consulta de ontologías (OLS, *Ontology Lookup Service*) (Cote, Jones et al. 2006).

Flujos de trabajo "olvidados"

Todos los repositorios componentes de ProteomeXchange trabajan, o soportan de manera más o menos automática, los datos de identificación de péptidos y proteínas basados en espectrometría de masas MS/MS en tándem con cromatografía líquida. Sin embargo, existen otros tipos de datos que o bien no están soportados o únicamente de manera superficial. Es el caso por ejemplo, de los datos provenientes de un experimento basado en geles. Un flujo de trabajo típico en proteómica se basa en el análisis de cada una de las manchas cortadas de un gel 2D en una placa de MALDI de un espectrómetro como un MALDI TOF TOF, en el que en cada pocillo se analiza una de las manchas que normalmente proviene de una única proteína. En este caso, los flujos de trabajo existentes no permiten el manejo fácil e intuitivo de dichos datos. Siendo cierto que la cantidad de los datos generados por estas otras técnicas no es tan grande como lo es en las técnicas de identificación a gran escala, existe una clara laguna para soportar estos flujos de trabajo, que no es comprensible teniendo en cuenta que los estándares

desarrollados por el HUPO-PSI soportan dichos datos. Es también el caso de los datos en sí de geles, con el estándar gelML (Gibson, Hoogland et al. 2010), o de los datos cromatográficos con el estándar spML (no publicado) que debido al poco interés de las casas comerciales en desarrollar herramientas informáticas que los soporten, no han prosperado y han quedado en el olvido. En nuestro laboratorio hemos intentado en cierta manera cubrir este vacío, creando un repositorio de experimentos proteómicos, incluyendo también los experimentos basados en geles (GE y GI) e implementando un sistema para exportar los datos de un documento MIAPE GE en un fichero gelML, el *gelML exporter*, o con la herramienta *PRIDE Spot Mapper* (http://proteo.cnb.csic.es/pridespotmapper/) que resuelve el problema de la creación de un único fichero PRIDE XML para todas las identificaciones de un gel. Sin embargo, la publicación de estos otros tipos de datos siempre será más costosa que la publicación de datos de identificación a gran escala por espectrometría de masas en tándem. A este respecto, y con la intención de mostrar los recursos bioinformáticos existentes para la publicación de datos provenientes de experimentos proteómicos basados en geles, se publicó un trabajo (Kenyani, Medina-Aunon et al. 2011) como ejemplo de las mejores prácticas para publicar y compartir datos de un experimento basado en DIGE (*Difference in-gel electrophoresis*) (Unlu, Morgan et al. 1997). En este flujo de trabajo se describía el experimento siguiendo las directrices MIAPE y usando nuestra herramienta online generadora de documentos MIAPE, tanto para la parte de separación por geles como para la parte de espectrometría de masas. Luego, se generaban dos ficheros estándares: uno el gelML (Gibson, Hoogland et al. 2010) con la información de todo el proceso de creación del gel DIGE y su análisis de imagen, y otro, el PRIDE XML con todos los datos de identificaciones por espectrometría de masas de los spots seleccionados como diferenciales, ambos con toda la información MIAPE correspondiente a los módulos GE, GI, MS y MSI. El flujo de creación de estos ficheros comprendía el uso de varias herramientas desarrolladas por nosotros.

Perspectivas de futuro

Los últimos estándares desarrollados por el HUPO-PSI son los relacionados con la cuantificación en Proteómica. El mzQuantML (Walzer, Qi et al. 2013) y el módulo MIAPE Quant (Martinez-Bartolome, Deutsch et al. 2013) han sido desarrollados en el contexto del HUPO-PSI con el objetivo de proporcionar estándares a una parte de la Proteómica que ha cobrado un gran protagonismo en los últimos años, en parte gracias a los avances tecnológicos que han mejorado la sensibilidad y reproducibilidad de estas técnicas. Parte del trabajo descrito en esta tesis comprende la definición de las directrices MIAPE para experimentos de Proteómica cuantitativa (MIAPE Quant), que dentro del amparo de la iniciativa HUPO-PSI, se realizó gracias a la coordinación de un grupo de expertos de ProteoRed con experiencia en diversas técnicas cuantitativas. Sin embargo, son todavía escasas las herramientas existentes

capaces de utilizar estos estándares y únicamente existe una versión preliminar de una librería en Java para el manejo de ficheros mzQuantML (jmzQuantML, http://code.google.com/p/jmzquantml/) y un validador semántico (http://code.google.com/p/mzquantml-validator/). Es de esperar por tanto que nuestro flujo de trabajo y nuestro repositorio de experimentos proteómicos soporten dichos estándares, para así contener y manejar también datos proteómicos cuantitativos, ya que sin duda los ficheros mzQuantML y las directrices MIAPE para cuantificación tendrán una gran utilidad en los próximos años.

Otras perspectivas de futuro se abren con respecto a los datos almacenados en el repositorio de experimentos MIAPE de ProteoRed, que gracias a la automatización proporcionada por la herramienta MIAPE Extractor, los datos contenidos en ella se han disparado en los últimos meses, llegando a tener más de 8.091.000 proteínas, 5.700.000 de ellas diferentes (o con distinto código de acceso) y más de 35.952.000 péptidos de los cuales 23.950.000 son secuencias peptídicas diferentes (Figura 55), prácticamente en un año de uso.

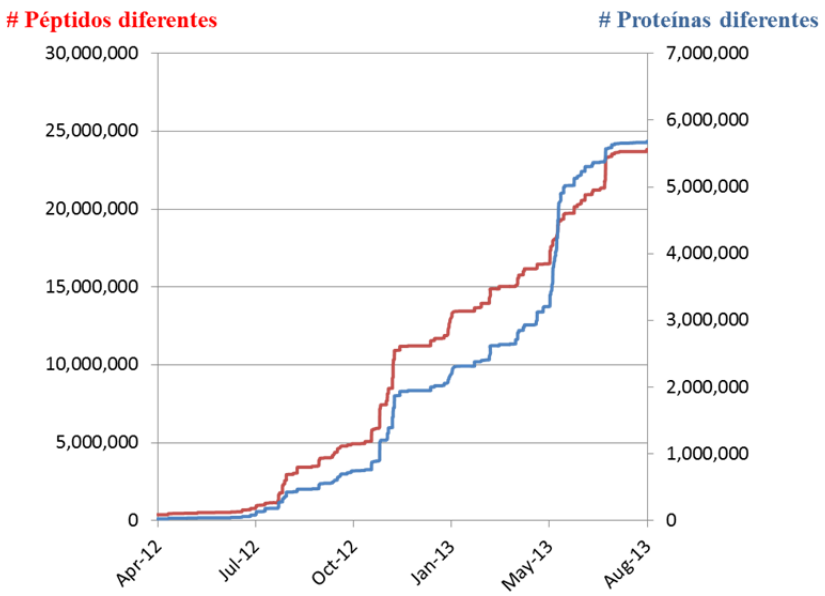


Figura 55. Aumento del número de péptidos y proteínas distintas a lo largo del tiempo en el repositorio de documentos MIAPE de ProteoRed: Se muestra el número de péptidos (línea roja) y proteínas (línea azul) distintos a lo largo del tiempo desde Abril de 2012 y Agosto de 2013.

Así pues, queda pendiente un trabajo de extracción de información (*data mining*) de los datos allí almacenados, para construir un portal de acceso y consulta de los datos almacenados, parecido a lo existente en PRIDE-Q (ej:

http://wwwdev.ebi.ac.uk/pride/prideq/viewer/#PROTEIN=A0A183). La intención es poder navegar de una forma intuitiva y fácil por los datos del repositorio, pudiendo buscar proteínas o secuencias peptídicas concretas, viendo los PSMs asociados a cada proteína y visualizando los espectros debidamente anotados. Para ello, será necesario realizar un gran trabajo de optimización de acceso a la información, indexando los millones de proteínas, péptidos y espectros almacenados en la base de datos y en los ficheros de entrada de creación de los MIAPE MS.

Conclusiones

6. Conclusiones

1) No puede establecerse un criterio estadístico universal basado en determinados umbrales de corte sobre la mejor o la segunda mejor puntuación, porque la distribución de dichas puntuaciones siempre va a depender de la distribución de calidades, y por tanto, del conjunto de espectros con el que estemos trabajando.

2) El método de validación de identificaciones a gran escala de la razón de probabilidades constituye por su sencillez, robustez, ausencia de parámetros ajustables y de funciones empíricas y mayor sensibilidad que otros métodos, un método particularmente idóneo para la automatización del proceso de identificación de péptidos en experimentos de identificación masiva de péptidos mediante espectrometría de masas.

3) Por su parte el método de validación de la calidad única constituye un método simple y directo que permite evaluar cada uno de los espectros independientemente por lo que no es necesario construir distribuciones de puntuaciones promedio a partir de los resultados globales de experimento.

4) Ambos métodos sientan las bases para el desarrollo de una puntuación normalizada que permita comparar los resultados obtenidos por diferentes algoritmos con condiciones de búsqueda distintas (Martinez-Bartolome, Navarro et al. 2008).

5) Los validadores de información semántica y MIAPE para los ficheros de representación de datos estándares mzML y mzIdentML proporcionan un recurso de gran utilidad para desarrolladores de nuevos software de creación y manejo de dichos estándares (Ghali, Krishna et al. 2013).

6) El repositorio de experimentos proteómicos basados en las directrices MIAPE y la herramienta de generación de documentos MIAPE constituyen el primer y único recurso bioinformático existente dedicado a las directrices MIAPE (Martinez-Bartolome, Blanco et al. 2010, Martinez-Bartolome, Medina-Aunon et al. 2010).

7) La automatización de la extracción de información MIAPE de ficheros de datos proteómicos es posible gracias a la librería ProteoRed Java MIAPE API, y la interacción programática con el repositorio de documentos MIAPE de ProteoRed es

posible gracias a los servicios web que permiten un acceso remoto y seguro (Medina-Aunon, Martinez-Bartolome et al. 2011).

8) La herramienta ProteoRed MIAPE Extractor permite un flujo de manejo y análisis de datos completo. De manera extraordinariamente potente e intuitiva, permite la integración de grandes cantidades de datos de identificaciones, así como su inspección y filtrado. Así mismo, es la primera herramienta que permite el envío de datos a ProteomeXchange asegurando el sello de calidad que proporcionan las directrices MIAPE (Medina-Aunon, Martinez-Bartolome et al. 2011, Vizcaino, Deutsch et al. 2013).

9) La herramienta ProteoRed MIAPE Extractor está siendo utilizada por el consorcio español del Proyecto del Proteoma Humano, para la recopilación de los datos de identificación a gran escala de los participantes y para el envío de los resultados obtenidos al repositorio ProteomeXchange (Segura, Medina-Aunon et al. 2013, Segura, Medina-Aunon et al. 2013).

10) El análisis centralizado de datos provenientes de diferentes plataformas permite homogeneizar las diferencias debidas al uso de múltiples flujos de trabajo y de análisis. La herramienta ProteoRed MIAPE Extractor es de gran utilidad para integrar y comparar resultados en un experimento multicentro, tanto para los datos analizados por diferentes flujos de análisis como para los datos reanalizados de forma centralizada.

Bibliografía

7. Bibliografía

1. Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. Encyclopedia of Measurement and Statistics. N. J. Salkind. Thousand Oaks, CA, Sage.
2. Aebersold, R. and D. R. Goodlett (2001). "Mass spectrometry in proteomics." Chem Rev 101(2): 269-295.
3. Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res 25(17): 3389-3402.
4. Anderson, D. C., W. Li, D. G. Payan and W. S. Noble (2003). "A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores." J Proteome Res 2(2): 137-146.
5. Armesilla, A. L., E. Lorenzo, P. Gomez del Arco, S. Martinez-Martinez, A. Alfranca and J. M. Redondo (1999). "Vascular endothelial growth factor activates nuclear factor of activated T cells in human endothelial cells: a role for tissue factor gene expression." Mol Cell Biol 19(3): 2032-2043.
6. Aston, F. W. (1918). "A positive ray spectrograph." Physical Review XI: 707-714.
7. Aston, F. W. (1919). "The mass-spectra of chemical elements." Phil. Mag. 39: 611-619.
8. Aston, F. W. (1920). "Isotopes and atomic weights." Nature 105(2646): 617-619.
9. Bafna, V. and N. Edwards (2001). "SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database." Bioinformatics 17 Suppl 1: S13-21.
10. Bald, T., J. Barth, A. Niehues, M. Specht, M. Hippler and C. Fufezan (2012). "pymzML--Python module for high-throughput bioinformatics on mass spectrometry data." Bioinformatics 28(7): 1052-1053.
11. Baldwin, M. A. (2004). "Protein identification by mass spectrometry: issues to be considered." Mol Cell Proteomics 3(1): 1-9.
12. Barber, M., R. S. Bordoli, R. D. Sedgwick, A. N. Tyler and B. W. Bycroft (1981). "Fast atom bombardment mass spectrometry of bleomycin A2 and B2 and their metal complexes." Biochem Biophys Res Commun 101(2): 632-638.
13. Barsnes, H., J. A. Vizcaino, I. Eidhammer and L. Martens (2009). "PRIDE Converter: making proteomics data-sharing easy." Nat Biotechnol 27(7): 598-599.
14. Barsnes, H., J. A. Vizcaino, F. Reisinger, I. Eidhammer and L. Martens (2011). "Submitting proteomics data to PRIDE using PRIDE Converter." Methods Mol Biol 694: 237-253.
15. Beavis, R. C. (2006). "Using the global proteome machine for protein identification." Methods Mol Biol 328: 217-228.
16. Beck, M., A. Schmidt, J. Malmstroem, M. Claassen, A. Ori, A. Szymborska, F. Herzog, O. Rinner, J. Ellenberg and R. Aebersold (2011). "The quantitative proteome of a human cell line." Mol Syst Biol 7: 549.
17. Bell, A. W., E. W. Deutsch, C. E. Au, R. E. Kearney, R. Beavis, S. Sechi, T. Nilsson, J. J. Bergeron and H. T. S. W. Group (2009). "A HUPO test sample study reveals common problems in mass spectrometry-based proteomics." Nat Methods 6(6): 423-430.

18. Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal R Stat Soc Ser B 57(1): 289-300.
19. Bern, M., D. Goldberg, W. H. McDonald and J. R. Yates, 3rd (2004). "Automatic quality assessment of peptide tandem mass spectra." Bioinformatics 20 Suppl 1: i49-54.
20. Binz, P. A., R. Barkovich, R. C. Beavis, D. Creasy, D. M. Horn, R. K. Julian, Jr., S. L. Seymour, C. F. Taylor and Y. Vandenbrouck (2008). "Guidelines for reporting the use of mass spectrometry informatics in proteomics." Nat Biotechnol 26(8): 862.
21. Bjellqvist, B., K. Ek, P. G. Righetti, E. Gianazza, A. Gorg, R. Westermeier and W. Postel (1982). "Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications." J Biochem Biophys Methods 6(4): 317-339.
22. Bossuyt, P. M., J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. M. Irwig, D. Moher, D. Rennie, H. C. de Vet, J. G. Lijmer and A. Standards for Reporting of Diagnostic (2003). "The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration." Clin Chem 49(1): 7-18.
23. Bourbeillon, J., S. Orchard, I. Benhar, C. Borrebaeck, A. de Daruvar, S. Dubel, R. Frank, F. Gibson, D. Gloriam, N. Haslam, T. Hiltker, I. Humphrey-Smith, M. Hust, D. Juncker, M. Koegl, Z. Konthur, B. Korn, S. Krobitsch, S. Muyldermans, P. A. Nygren, S. Palcy, B. Polic, H. Rodriguez, A. Sawyer, M. Schlapshy, M. Snyder, O. Stoevesandt, M. J. Taussig, M. Templin, M. Uhlen, S. van der Maarel, C. Wingren, H. Hermjakob and D. Sherman (2010). "Minimum information about a protein affinity reagent (MIAPAR)." Nat Biotechnol 28(7): 650-653.
24. Bradshaw, R. A., A. L. Burlingame, S. Carr and R. Aebersold (2006). "Reporting protein identification data: the next generation of guidelines." Mol Cell Proteomics 5(5): 787-788.
25. Brazma, A., P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo and M. Vingron (2001). "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data." Nat Genet 29(4): 365-371.
26. Carr, S., R. Aebersold, M. Baldwin, A. Burlingame, K. Clauser, A. Nesvizhskii, P. Working Group on Publication Guidelines for and D. Protein Identification (2004). "The need for guidelines in publication of peptide and protein identification data: Working Group on Publication Guidelines for Peptide and Protein Identification Data." Mol Cell Proteomics 3(6): 531-533.
27. Celis, J. E. (2004). "Gel-based proteomics: what does MCP expect?" Mol Cell Proteomics 3(10): 949.
28. Chambers, M. C., B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M. Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb and P. Mallick (2012). "A cross-platform toolkit for mass spectrometry and proteomics." Nat Biotechnol 30(10): 918-920.
29. Choi, H., D. Ghosh and A. I. Nesvizhskii (2008). "Statistical validation of peptide identifications in large-scale proteomics

using the target-decoy database search strategy and flexible mixture modeling." J Proteome Res 7(1): 286-292.

30. Choi, H. and A. I. Nesvizhskii (2008). "False discovery rates and related statistical concepts in mass spectrometry-based proteomics." J Proteome Res 7(1): 47-50.
31. Colinge, J., A. Masselot, I. Cusin, E. Mahe, A. Niknejad, G. Argoud-Puy, S. Reffas, N. Bederr, A. Gleizes, P. A. Rey and L. Bougueleret (2004). "High-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics." Proteomics 4(7): 1977-1984.
32. Colinge, J., A. Masselot, M. Giron, T. Dessingy and J. Magnin (2003). "OLAV: towards high-throughput tandem mass spectrometry data identification." Proteomics 3(8): 1454-1463.
33. Comisarow, M. B. and A. G. Marshall (1974). "Fourier transform ion cyclotron resonance spectroscopy." Chemical Physics Letters 25: 282-283.
34. Comisarow, M. B. and A. G. Marshall (1996). "The early development of Fourier transform ion cyclotron resonance (FT-ICR) spectroscopy." J Mass Spectrom 31(6): 581-585.
35. Cote, R. G., J. Griss, J. A. Dianes, R. Wang, J. C. Wright, H. W. van den Toorn, B. van Breukelen, A. J. Heck, N. Hulstaert, L. Martens, F. Reisinger, A. Csordas, D. Ovelleiro, Y. Perez-Rivevol, H. Barsnes, H. Hermjakob and J. A. Vizcaino (2012). "The PRoteomics IDEntification (PRIDE) Converter 2 framework: an improved suite of tools to facilitate data submission to the PRIDE database and the ProteomeXchange consortium." Mol Cell Proteomics 11(12): 1682-1689.
36. Cote, R. G., P. Jones, R. Apweiler and H. Hermjakob (2006). "The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries." BMC Bioinformatics 7: 97.
37. Cote, R. G., F. Reisinger and L. Martens (2010). "jmzML, an open-source Java API for mzML, the PSI standard for MS data." Proteomics 10(7): 1332-1335.
38. Craig, R. and R. C. Beavis (2004). "TANDEM: matching proteins with tandem mass spectra." Bioinformatics 20(9): 1466-1467.
39. Csordas, A., R. Wang, D. Rios, F. Reisinger, J. M. Foster, D. J. Slotta, J. A. Vizcaino and H. Hermjakob (2013). "From Peptidome to PRIDE: Public proteomics data migration at a large scale." Proteomics.
40. de Godoy, L. M., J. V. Olsen, J. Cox, M. L. Nielsen, N. C. Hubner, F. Frohlich, T. C. Walther and M. Mann (2008). "Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast." Nature 455(7217): 1251-1254.
41. Dempster, A. J. (1917). "A new method os positive ray analysis." Physical Review XI(4): 316-325.
42. Desiere, F., E. W. Deutsch, N. L. King, A. I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S. N. Loevenich and R. Aebersold (2006). "The PeptideAtlas project." Nucleic Acids Res 34(Database issue): D655-658.
43. Desiere, F., E. W. Deutsch, A. I. Nesvizhskii, P. Mallick, N. L. King, J. K. Eng, A. Aderem, R. Boyle, E. Brunner, S. Donohoe, N. Fausto, E. Hafen, L. Hood, M. G. Katze, K. A. Kennedy, F. Kregenow, H. Lee, B. Lin, D. Martin, J. A. Ranish, D. J. Rawlings, L. E. Samelson, Y. Shiio, J. D. Watts, B. Wollscheid, M. E. Wright, W. Yan, L. Yang, E. C. Yi, H. Zhang and R. Aebersold (2005). "Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry." Genome Biol 6(1): R9.

44. Deutsch, E. (2008). "mzML: a single, unifying data format for mass spectrometer output." Proteomics 8(14): 2776-2777.
45. Deutsch, E. W. (2010). "Mass spectrometer output file format mzML." Methods Mol Biol 604: 319-331.
46. Deutsch, E. W. (2010). "The PeptideAtlas Project." Methods Mol Biol 604: 285-296.
47. Deutsch, E. W., M. Chambers, S. Neumann, F. Levander, P. A. Binz, J. Shofstahl, D. S. Campbell, L. Mendoza, D. Ovelleiro, K. Helsens, L. Martens, R. Aebersold, R. L. Moritz and M. Y. Brusniak (2012). "TraML--a standard format for exchange of selected reaction monitoring transition lists." Mol Cell Proteomics 11(4): R111 015040.
48. Domann, P. J., S. Akashi, C. Barbas, L. Huang, W. Lau, C. Legido-Quigley, S. McClean, C. Neususs, D. Perrett, M. Quaglia, E. Rapp, L. Smallshaw, N. W. Smith, W. F. Smyth, C. F. Taylor and E. Minimum Information About a Proteomics (2010). "Guidelines for reporting the use of capillary electrophoresis in proteomics." Nat Biotechnol 28(7): 654-655.
49. Eisenacher, M. (2011). "mzIdentML: an open community-built standard format for the results of proteomics spectrum identification algorithms." Methods Mol Biol 696: 161-177.
50. Elias, J. E. and S. P. Gygi (2007). "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry." Nat Methods 4(3): 207-214.
51. Emmett, M. R. and R. M. Caprioli (1994). "Micro-electrospray mass spectrometry: ultra-high-sensitivity analysis of peptides and proteins." Journal of the American Society for Mass Spectrometry 5: 605-613.
52. Eng, J., A. L. McCormack and J. R. Yates (1994). "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database." Journal of the American Society for Mass Spectrometry 5: 976-989.
53. Eng, J. K., B. Fischer, J. Grossmann and M. J. Maccoss (2008). "A fast SEQUEST cross correlation algorithm." J Proteome Res 7(10): 4598-4602.
54. Farrah, T., E. W. Deutsch and R. Aebersold (2011). "Using the Human Plasma PeptideAtlas to study human plasma proteins." Methods Mol Biol 728: 349-374.
55. Fenn, J. B., M. Mann, C. K. Meng, S. F. Wong and C. M. Whitehouse (1989). "Electrospray ionization for mass spectrometry of large biomolecules." Science 246(4926): 64-71.
56. Fenyo, D. and R. C. Beavis (2003). "A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes." Anal Chem 75(4): 768-774.
57. Fenyo, D., J. Eriksson and R. Beavis (2010). "Mass spectrometric protein identification using the global proteome machine." Methods Mol Biol 673: 189-202.
58. Florens, L., M. P. Washburn, J. D. Raine, R. M. Anthony, M. Grainger, J. D. Haynes, J. K. Moch, N. Muster, J. B. Sacci, D. L. Tabb, A. A. Witney, D. Wolters, Y. Wu, M. J. Gardner, A. A. Holder, R. E. Sinden, J. R. Yates and D. J. Carucci (2002). "A proteomic view of the Plasmodium falciparum life cycle." Nature 419(6906): 520-526.
59. Geer, L. Y., S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi and S. H. Bryant (2004). "Open mass spectrometry search algorithm." J Proteome Res 3(5): 958-964.
60. Ghali, F., R. Krishna, P. Lukasse, S. Martinez-Bartolome, F. Reisinger, H. Hermjakob, J. A. Vizcaino and A. R. Jones (2013). "A toolkit for mzIdentML: the ProteoIDViewer, the mzidLibrary and the mzidValidator." Molecular & Cellular Proteomics (Under review).

61. Ghali, F., R. Krishna, P. Lukasse, S. Martinez-Bartolome, F. Reisinger, H. Hermjakob, J. A. Vizcaino and A. R. Jones (2013). "A toolkit for the mzIdentML standard: the ProteoIDViewer, the mzidLibrary and the mzidValidator." Mol Cell Proteomics.
62. Gharahdaghi, F., C. R. Weinberg, D. A. Meagher, B. S. Imai and S. M. Mische (1999). "Mass spectrometric identification of proteins from silver-stained polyacrylamide gel: a method for the removal of silver ions to enhance sensitivity." Electrophoresis 20(3): 601-605.
63. Gibson, F., L. Anderson, G. Babnigg, M. Baker, M. Berth, P. A. Binz, A. Borthwick, P. Cash, B. W. Day, D. B. Friedman, D. Garland, H. B. Gutstein, C. Hoogland, N. A. Jones, A. Khan, J. Klose, A. I. Lamond, P. F. Lemkin, K. S. Lilley, J. Minden, N. J. Morris, N. W. Paton, M. R. Pisano, J. E. Prime, T. Rabilloud, D. A. Stead, C. F. Taylor, H. Voshol, A. Wipat and A. R. Jones (2008). "Guidelines for reporting the use of gel electrophoresis in proteomics." Nat Biotechnol 26(8): 863-864.
64. Gibson, F., C. Hoogland, S. Martinez-Bartolome, J. A. Medina-Aunon, J. P. Albar, G. Babnigg, A. Wipat, H. Hermjakob, J. S. Almeida, R. Stanislaus, N. W. Paton and A. R. Jones (2010). "The gel electrophoresis markup language (GelML) from the Proteomics Standards Initiative." Proteomics 10(17): 3073-3081.
65. Gorg, A., W. Postel and S. Gunther (1988). "The current state of two-dimensional electrophoresis with immobilized pH gradients." Electrophoresis 9(9): 531-546.
66. Gygi, S. P., G. L. Corthals, Y. Zhang, Y. Rochon and R. Aebersold (2000). "Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology." Proc Natl Acad Sci U S A 97(17): 9390-9395.
67. Henzel, W. J., T. M. Billeci, J. T. Stults, S. C. Wong, C. Grimley and C. Watanabe (1993). "Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases." Proc Natl Acad Sci U S A 90(11): 5011-5015.
68. Hermjakob, H. and R. Apweiler (2006). "The Proteomics Identifications Database (PRIDE) and the ProteomExchange Consortium: making proteomics data accessible." Expert Rev Proteomics 3(1): 1-3.
69. Hill, J. A., B. E. Smith, P. G. Papoulias and P. C. Andrews (2010). "ProteomeCommons.org collaborative annotation and project management resource integrated with the Tranche repository." J Proteome Res 9(6): 2809-2811.
70. Hoogland, C., M. O'Gorman, P. Bogard, F. Gibson, M. Berth, S. J. Cockell, A. Ekefjard, O. Forsstrom-Olsson, A. Kapferer, M. Nilsson, S. Martinez-Bartolome, J. P. Albar, S. Echevarria-Zomeno, M. Martinez-Gomariz, J. Joets, P. A. Binz, C. F. Taylor, A. Dowsey, A. R. Jones and E. Minimum Information About a Proteomics (2010). "Guidelines for reporting the use of gel image informatics in proteomics." Nat Biotechnol 28(7): 655-656.
71. Horth, P., C. A. Miller, T. Preckel and C. Wenz (2006). "Efficient fractionation and improved protein identification by peptide OFFGEL electrophoresis." Mol Cell Proteomics 5(10): 1968-1974.
72. Hu, Q., R. J. Noll, H. Li, A. Makarov, M. Hardman and R. Graham Cooks (2005). "The Orbitrap: a new mass spectrometer." J Mass Spectrom 40(4): 430-443.
73. Huttenhain, R., S. Surinova, R. Ossola, Z. Sun, D. Campbell, F. Cerciello, R. Schiess, D. Bausch-Fluck, G. Rosenberger, J. Chen, O. Rinner, U. Kusebauch, M. Hajduch, R. L. Moritz, B. Wollscheid and R. Aebersold (2013). "N-Glycoprotein SRMAtlas: A

Resource of mass spectrometric assays for n-glycosites enabling consistent and multiplexed protein quantification for clinical applications." Mol Cell Proteomics 12(4): 1005-1016.

74. James, P., M. Quadroni, E. Carafoli and G. Gonnet (1993). "Protein identification by mass profile fingerprinting." Biochem Biophys Res Commun 195(1): 58-64.
75. Ji, L., T. Barrett, O. Ayanbule, D. B. Troup, D. Rudnev, R. N. Muertter, M. Tomashevsky, A. Soboleva and D. J. Slotta (2010). "NCBI Peptidome: a new repository for mass spectrometry proteomics data." Nucleic Acids Res 38(Database issue): D731-735.
76. Johnson, R. S., S. A. Martin, K. Biemann, J. T. Stults and J. T. Watson (1987). "Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine." Anal Chem 59(21): 2621-2625.
77. Johnson, R. S. and A. O. Nier (1953). "Angular aberrations in sector shaped electromagnetic lenses for focusing beams of charged peptides." Physical Review 91(1): 10-17.
78. Jones, A. R., K. Carroll, D. Knight, K. Maclellan, P. J. Domann, C. Legido-Quigley, L. Huang, L. Smallshaw, H. Mirzaei, J. Shofstahl, N. W. Paton and E. Minimum Information About a Proteomics (2010). "Guidelines for reporting the use of column chromatography in proteomics." Nat Biotechnol 28(7): 654.
79. Jones, A. R., M. Eisenacher, G. Mayer, O. Kohlbacher, J. Siepen, S. J. Hubbard, J. N. Selley, B. C. Searle, J. Shofstahl, S. L. Seymour, R. Julian, P. A. Binz, E. W. Deutsch, H. Hermjakob, F. Reisinger, J. Griss, J. A. Vizcaino, M. Chambers, A. Pizarro and D. Creasy (2012). "The mzIdentML data standard for mass spectrometry-based proteomics results." Mol Cell Proteomics 11(7): M111 014381.
80. Jones, P. and L. Martens (2010). "Using the PRIDE proteomics identifications database for knowledge discovery and data analysis." Methods Mol Biol 604: 297-307.
81. Käll, L., J. D. Storey, M. J. MacCoss and W. S. Noble (2008). "Assigning significance to peptides identified by tandem mass spectrometry using decoy databases." J Proteome Res 7(1): 29-34.
82. Käll, L., J. D. Storey, M. J. MacCoss and W. S. Noble (2008). "Posterior error probabilities and false discovery rates: two sides of the same coin." J Proteome Res 7(1): 40-44.
83. Käll, L., J. D. Storey and W. S. Noble (2008). "Non-parametric estimation of posterior error probabilities associated with peptides identified by tandem mass spectrometry." Bioinformatics 24(16): i42-48.
84. Karas, M., D. Bachmann and F. Hillenkamp (1985). "Influence of the Wavelength in High-Irradiance Ultraviolet Laser Desorption Mass Spectrometry of Organic Molecules." Analytical Chemistry 57(14): 2935-2939.
85. Keller, A., A. I. Nesvizhskii, E. Kolker and R. Aebersold (2002). "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search." Anal Chem 74(20): 5383-5392.
86. Keller, A. and D. Shteynberg (2011). "Software pipeline and data analysis for MS/MS proteomics: the trans-proteomic pipeline." Methods Mol Biol 694: 169-189.
87. Kenyani, J., J. A. Medina-Aunon, S. Martinez-Bartolome, J. P. Albar, J. M. Wastling and A. R. Jones (2011). "A DIGE study on the effects of salbutamol on the rat muscle proteome - an exemplar of best practice for data sharing in proteomics." BMC Res Notes 4: 86.

88. Kessner, D., M. Chambers, R. Burke, D. Agus and P. Mallick (2008). "ProteoWizard: open source software for rapid proteomics tools development." Bioinformatics 24(21): 2534-2536.
89. Kettner, C., D. Field, S. A. Sansone, C. Taylor, J. Aerts, N. Binns, A. Blake, C. M. Britten, A. de Marco, J. Fostel, P. Gaudet, A. Gonzalez-Beltran, N. Hardy, J. Hellemans, H. Hermjakob, N. Juty, J. Leebens-Mack, E. Maguire, S. Neumann, S. Orchard, H. Parkinson, W. Piel, S. Ranganathan, P. Rocca-Serra, A. Santarsiero, D. Shotton, P. Sterk, A. Untergasser and P. L. Whetzel (2010). "Meeting Report from the Second "Minimum Information for Biological and Biomedical Investigations" (MIBBI) workshop." Stand Genomic Sci 3(3): 259-266.
90. Kim, S., N. Gupta and P. A. Pevzner (2008). "Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases." J Proteome Res 7(8): 3354-3363.
91. Kislinger, T., K. Rahman, D. Radulovic, B. Cox, J. Rossant and A. Emili (2003). "PRISM, a generic large scale proteomic investigation strategy for mammals." Mol Cell Proteomics 2(2): 96-106.
92. Klose, J. (1975). "Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals." Humangenetik 26(3): 231-243.
93. Laemmli, U. K. (1970). "Cleavage of structural proteins during the assembly of the head of bacteriophage T4." Nature 227(5259): 680-685.
94. Link, A. J., J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, D. R. Morris, B. M. Garvik and J. R. Yates, 3rd (1999). "Direct analysis of protein complexes using mass spectrometry." Nat Biotechnol 17(7): 676-682.
95. Lopez-Ferrer, D., S. Martinez-Bartolome, M. Villar, M. Campillos, F. Martin-Maroto and J. Vazquez (2004). "Statistical model for large-scale peptide identification in databases from tandem mass spectra using SEQUEST." Anal Chem 76(23): 6853-6860.
96. Lundby, A. and J. V. Olsen (2011). "GeLCMS for in-depth protein characterization and advanced analysis of proteomes." Methods Mol Biol 753: 143-155.
97. MacCoss, M. J., C. C. Wu and J. R. Yates, 3rd (2002). "Probability-based validation of protein identifications using a modified SEQUEST algorithm." Anal Chem 74(21): 5593-5599.
98. Macfarlane, R. D. (1990). "Principles of californium-252 plasma desorption mass spectrometry applied to protein analysis." Methods Enzymol 193: 263-280.
99. Magnin, J., A. Masselot, C. Menzel and J. Colinge (2004). "OLAV-PMF: a novel scoring scheme for high-throughput peptide mass fingerprinting." J Proteome Res 3(1): 55-60.
100. Makarov, A. (2000). "Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis." Anal Chem 72(6): 1156-1162.
101. Mann, M., P. Hojrup and P. Roepstorff (1993). "Use of mass spectrometric molecular weight information to identify proteins in sequence databases." Biol Mass Spectrom 22(6): 338-345.
102. Martens, L. (2013). "Resilience in the proteomics data ecosystem: how the field cares for its data." Proteomics.
103. Martens, L., M. Chambers, M. Sturm, D. Kessner, F. Levander, J. Shofstahl, W. H. Tang, A. Rompp, S. Neumann, A. D. Pizarro, L. Montecchi-Palazzi, N. Tasman, M. Coleman, F. Reisinger, P. Souda, H. Hermjakob, P. A. Binz and E. W. Deutsch

(2011). "mzML--a community standard for mass spectrometry data." Mol Cell Proteomics 10(1): R110 000133.

104. Martens, L., H. Hermjakob, P. Jones, M. Adamski, C. Taylor, D. States, K. Gevaert, J. Vandekerckhove and R. Apweiler (2005). "PRIDE: the proteomics identifications database." Proteomics 5(13): 3537-3545.
105. Martinez-Bartolome, S., F. Blanco and J. P. Albar (2010). "Relevance of proteomics standards for the ProteoRed Spanish organization." J Proteomics 73(6): 1061-1066.
106. Martinez-Bartolome, S., E. W. Deutsch, P. A. Binz, A. R. Jones, M. Eisenacher, G. Mayer, A. Campos, F. Canals, J. J. Bech-Serra, M. Carrascal, M. Gay, A. Paradela, R. Navajas, M. Marcilla, M. L. Hernaez, M. D. Gutierrez-Blazquez, L. F. Velarde, K. Aloria, J. Beaskoetxea, J. A. Medina-Aunon and J. P. Albar (2013). "Guidelines for reporting quantitative mass spectrometry based experiments in proteomics." J Proteomics.
107. Martinez-Bartolome, S., J. A. Medina-Aunon, A. R. Jones and J. P. Albar (2010). "Semi-automatic tool to describe, store and compare proteomics experiments based on MIAPE compliant reports." Proteomics 10(6): 1256-1260.
108. Martinez-Bartolome, S., P. Navarro, F. Martin-Maroto, D. Lopez-Ferrer, A. Ramos-Fernandez, M. Villar, J. P. Garcia-Ruiz and J. Vazquez (2008). "Properties of average score distributions of SEQUEST: the probability ratio method." Mol Cell Proteomics 7(6): 1135-1145.
109. Medina-Aunon, J. A., S. Martinez-Bartolome, M. A. Lopez-Garcia, E. Salazar, R. Navajas, A. R. Jones, A. Paradela and J. P. Albar (2011). "The ProteoRed MIAPE web toolkit: a user-friendly framework to connect and share proteomics standards." Mol Cell Proteomics 10(10): M111 008334.
110. Meyer-Arendt, K., W. M. Old, S. Houel, K. Renganathan, B. Eichelberger, K. A. Resing and N. G. Ahn (2011). "IsoformResolver: A peptide-centric algorithm for protein inference." J Proteome Res 10(7): 3060-3075.
111. Montecchi-Palazzi, L., S. Kerrien, F. Reisinger, B. Aranda, A. R. Jones, L. Martens and H. Hermjakob (2009). "The PSI semantic validator: a framework to check MIAPE compliance of proteomics data." Proteomics 9(22): 5112-5119.
112. Moore, R. E., M. K. Young and T. D. Lee (2002). "Qscore: an algorithm for evaluating SEQUEST database search results." J Am Soc Mass Spectrom 13(4): 378-386.
113. Mortz, E., P. B. O'Connor, P. Roepstorff, N. L. Kelleher, T. D. Wood, F. W. McLafferty and M. Mann (1996). "Sequence tag identification of intact proteins by matching tanden mass spectral data against sequence data bases." Proc Natl Acad Sci U S A 93(16): 8264-8267.
114. Muth, T., M. Vaudel, H. Barsnes, L. Martens and A. Sickmann (2010). "XTandem Parser: an open-source library to parse and analyse X!Tandem MS/MS search results." Proteomics 10(7): 1522-1524.
115. Nagaraj, N., N. A. Kulak, J. Cox, N. Neuhauser, K. Mayr, O. Hoerning, O. Vorm and M. Mann (2012). "System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap." Mol Cell Proteomics 11(3): M111 013722.
116. Nagaraj, N., J. R. Wisniewski, T. Geiger, J. Cox, M. Kircher, J. Kelso, S. Paabo and M. Mann (2011). "Deep proteome and transcriptome mapping of a human cancer cell line." Mol Syst Biol 7: 548.
117. Navarro, P. and J. Vazquez (2009). "A refined method to calculate false discovery rates for peptide identification using decoy databases." J Proteome Res 8(4): 1792-1796.

118. Nesvizhskii, A. I. and R. Aebersold (2005). "Interpretation of shotgun proteomic data: the protein inference problem." Mol Cell Proteomics 4(10): 1419-1440.
119. Nesvizhskii, A. I., A. Keller, E. Kolker and R. Aebersold (2003). "A statistical model for identifying proteins by tandem mass spectrometry." Anal Chem 75(17): 4646-4658.
120. Nesvizhskii, A. I., F. F. Roos, J. Grossmann, M. Vogelzang, J. S. Eddes, W. Gruissem, S. Baginsky and R. Aebersold (2006). "Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides." Mol Cell Proteomics 5(4): 652-670.
121. Nesvizhskii, A. I., O. Vitek and R. Aebersold (2007). "Analysis and validation of proteomic data generated by tandem mass spectrometry." Nat Methods 4(10): 787-797.
122. O'Farrell, P. H. (1975). "High resolution two-dimensional electrophoresis of proteins." J Biol Chem 250(10): 4007-4021.
123. Ogueta, S., J. Munoz, E. Obregon, E. Delgado-Baeza and J. P. Garcia-Ruiz (2002). "Prolactin is a component of the human synovial liquid and modulates the growth and chondrogenic differentiation of bone marrow-derived mesenchymal stem cells." Mol Cell Endocrinol 190(1-2): 51-63.
124. Orchard, S., B. Al-Lazikani, S. Bryant, D. Clark, E. Calder, I. Dix, O. Engkvist, M. Forster, A. Gaulton, M. Gilson, R. Glen, M. Grigorov, K. Hammond-Kosack, L. Harland, A. Hopkins, C. Larminie, N. Lynch, R. K. Mann, P. Murray-Rust, E. Lo Piparo, C. Southan, C. Steinbeck, D. Wishart, H. Hermjakob, J. Overington and J. Thornton (2011). "Minimum information about a bioactive entity (MIABE)." Nat Rev Drug Discov 10(9): 661-669.
125. Orchard, S., P. A. Binz and H. Hermjakob (2009). "Second Joint HUPO publication and Proteomics Standards Initiative workshop." Proteomics 9(19): 4426-4428.
126. Orchard, S., H. Hermjakob, C. Taylor, P. A. Binz, C. Hoogland, R. Julian, J. S. Garavelli, R. Aebersold and R. Apweiler (2006). "Autumn 2005 Workshop of the Human Proteome Organisation Proteomics Standards Initiative (HUPO-PSI) Geneva, September, 4-6, 2005." Proteomics 6(3): 738-741.
127. Orchard, S., H. Hermjakob, C. F. Taylor, F. Potthast, P. Jones, W. Zhu, R. K. Julian, Jr. and R. Apweiler (2005). "Further steps in standardisation. Report of the second annual Proteomics Standards Initiative Spring Workshop (Siena, Italy 17-20th April 2005)." Proteomics 5(14): 3552-3555.
128. Orchard, S. and P. Ping (2009). "HUPO World Congress Publication Committee meeting. August 2008, Amsterdam, The Netherlands." Proteomics 9(3): 502-503.
129. Orchard, S., L. Salwinski, S. Kerrien, L. Montecchi-Palazzi, M. Oesterheld, V. Stumpflen, A. Ceol, A. Chatr-aryamontri, J. Armstrong, P. Woollard, J. J. Salama, S. Moore, J. Wojcik, G. D. Bader, M. Vidal, M. E. Cusick, M. Gerstein, A. C. Gavin, G. Superti-Furga, J. Greenblatt, J. Bader, P. Uetz, M. Tyers, P. Legrain, S. Fields, N. Mulder, M. Gilson, M. Niepmann, L. Burgoon, J. De Las Rivas, C. Prieto, V. M. Perreau, C. Hogue, H. W. Mewes, R. Apweiler, I. Xenarios, D. Eisenberg, G. Cesareni and H. Hermjakob (2007). "The minimum information required for reporting a molecular interaction experiment (MIMIx)." Nat Biotechnol 25(8): 894-898.
130. Orchard, S., C. Taylor, H. Hermjakob, W. Zhu, R. Julian and R. Apweiler (2004). "Current status of proteomic standards development." Expert Rev Proteomics 1(2): 179-183.

131. Pappin, D. J., P. Hojrup and A. J. Bleasby (1993). "Rapid identification of proteins by peptide-mass fingerprinting." Curr Biol 3(6): 327-332.
132. Paradela, A., P. R. Escuredo and J. P. Albar (2006). "Geographical focus. Proteomics initiatives in Spain: ProteoRed." Proteomics 6 Suppl 2: 73-76.
133. Paul, W. and H. Steinwedel (1953). "Ein neues Massenspecktrometer ohne magnetfeld." Z. Naturforschg. 8: 448-450.
134. Pedrioli, P. G., J. K. Eng, R. Hubley, M. Vogelzang, E. W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R. H. Angeletti, R. Apweiler, K. Cheung, C. E. Costello, H. Hermjakob, S. Huang, R. K. Julian, E. Kapp, M. E. McComb, S. G. Oliver, G. Omenn, N. W. Paton, R. Simpson, R. Smith, C. F. Taylor, W. Zhu and R. Aebersold (2004). "A common open representation of mass spectrometry data and its application to proteomics research." Nat Biotechnol 22(11): 1459-1466.
135. Peng, J., J. E. Elias, C. C. Thoreen, L. J. Licklider and S. P. Gygi (2003). "Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome." J Proteome Res 2(1): 43-50.
136. Picotti, P., M. Clement-Ziza, H. Lam, D. S. Campbell, A. Schmidt, E. W. Deutsch, H. Rost, Z. Sun, O. Rinner, L. Reiter, Q. Shen, J. J. Michaelson, A. Frei, S. Alberti, U. Kusebauch, B. Wollscheid, R. L. Moritz, A. Beyer and R. Aebersold (2013). "A complete mass-spectrometric map of the yeast proteome applied to quantitative trait analysis." Nature 494(7436): 266-270.
137. Picotti, P., H. Lam, D. Campbell, E. W. Deutsch, H. Mirzaei, J. Ranish, B. Domon and R. Aebersold (2008). "A database of mass spectrometric assays for the yeast proteome." Nat Methods 5(11): 913-914.
138. Prieto, G., K. Aloria, N. Osinalde, A. Fullaondo, J. M. Arizmendi and R. Matthiesen (2012). "PAnalyzer: a software tool for protein inference in shotgun proteomics." BMC Bioinformatics 13: 288.
139. Qeli, E. and C. H. Ahrens (2010). "PeptideClassifier for protein inference and targeted quantitative proteomics." Nat Biotechnol 28(7): 647-650.
140. Qian, W. J., T. Liu, M. E. Monroe, E. F. Strittmatter, J. M. Jacobs, L. J. Kangas, K. Petritis, D. G. Camp, 2nd and R. D. Smith (2005). "Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome." J Proteome Res 4(1): 53-62.
141. Razumovskaya, J., V. Olman, D. Xu, E. C. Uberbacher, N. C. VerBerkmoes, R. L. Hettich and Y. Xu (2004). "A computational method for assessing peptide- identification reliability in tandem mass spectrometry analysis with SEQUEST." Proteomics 4(4): 961-969.
142. Reimers, M. and V. J. Carey (2006). "Bioconductor: an open source framework for bioinformatics and computational biology." Methods Enzymol 411: 119-134.
143. Reisinger, F., R. Krishna, F. Ghali, D. Rios, H. Hermjakob, J. A. Vizcaino and A. R. Jones (2012). "jmzIdentML API: A Java interface to the mzIdentML standard for peptide and protein identification data." Proteomics 12(6): 790-794.
144. Robin, X., C. Hoogland, R. D. Appel and F. Lisacek (2008). "MIAPEGelDB, a web-based submission tool and public repository for MIAPE gel electrophoresis documents." J Proteomics 71(2): 249-251.
145. Roepstorff, P. and J. Fohlman (1984). "Proposal for a common nomenclature for sequence ions in mass spectra of peptides." Biomed Mass Spectrom 11(11): 601.
146. Roepstorff, P. and W. J. Richter (1992). "Status of, and developments in, mass

spectrometry of peptides and proteins." International Journal of Mass Spectrometry and Ion Processes 118/119: 789-809.

147. Sadygov, R. G. and J. R. Yates, 3rd (2003). "A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases." Anal Chem 75(15): 3792-3798.
148. Schagger, H., H. Aquila and G. Von Jagow (1988). "Coomassie blue-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for direct visualization of polypeptides during electrophoresis." Anal Biochem 173(1): 201-205.
149. Schagger, H. and G. von Jagow (1987). "Tricine-sodium dodecyl sulfate-polyacrylamide gel electrophoresis for the separation of proteins in the range from 1 to 100 kDa." Anal Biochem 166(2): 368-379.
150. Schwartz, J. C., M. W. Senko and J. E. Syka (2002). "A two-dimensional quadrupole ion trap mass spectrometer." J Am Soc Mass Spectrom 13(6): 659-669.
151. Segura, V., A. Medina-Aunon, M. Mora, S. Martínez-Bartolomé, J. Abian, K. Aloria, O. Antúnez, J. Arizmendi, M. Azkargorta, S. Barceló, J. Beaskoetxea, J. Bech-Serra, F. J. Blanco, M. Braga-Monteiro, D. Cáceres, F. Canals, M. Carrascal, J. I. Casal, F. Clemente, N. Colome, N. Dasilva, P. Díaz, F. Elortza, P. Fernández-Puente, M. Fuentes, O. Gallardo, S. Gharbi, C. Gil, M. Hernáez, M. Lombardia, M. Lopez-Lucendo, M. Marcilla, J. Mato, M. Mendes, E. Oliveira, I. Orera, A. Pascual, G. Prieto, C. Ruiz-Romero, M. Sánchez del Pino, D. Tabas-Madrid, M. Valero, V. Vialas, J. Villanueva, J. P. Albar and F. Corrales (2013). "Surfing transcriptomic landscapes. A step beyond the annotation of Chromosome 16 proteome." Journal of Proteome Research (submitted).
152. Segura, V., J. A. Medina-Aunon, E. Guruceaga, S. I. Gharbi, C. Gonzalez-Tejedo, M. M. Sanchez del Pino, F. Canals, M. Fuentes, J. I. Casal, S. Martinez-Bartolome, F. Elortza, J. M. Mato, J. M. Arizmendi, J. Abian, E. Oliveira, C. Gil, F. Vivanco, F. Blanco, J. P. Albar and F. J. Corrales (2013). "Spanish human proteome project: dissection of chromosome 16." J Proteome Res 12(1): 112-122.
153. Senko, M. W., C. L. Hendrickson, L. Pasa-Tolic, J. A. Marto, F. M. White, S. Guan and A. G. Marshall (1996). "Electrospray ionization Fourier transform ion cyclotron resonance at 9.4 T." Rapid Commun Mass Spectrom 10(14): 1824-1828.
154. Slotta, D. J., T. Barrett and R. Edgar (2009). "NCBI Peptidome: a new public repository for mass spectrometry peptide identifications." Nat Biotechnol 27(7): 600-601.
155. Smedley, D., S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson and A. Kasprzyk (2009). "BioMart--biological queries made easy." BMC Genomics 10: 22.
156. Smith, B. E., J. A. Hill, M. A. Gjukich and P. C. Andrews (2011). "Tranche distributed repository and ProteomeCommons.org." Methods Mol Biol 696: 123-145.
157. Soric, B. (1989). "Statistical discoveries and effect-size estimation." J. Am. Stat. Assoc 84: 6008-6610.
158. Stephens, W. E. (1946). "Proceedings of the American Physica Society. J1. A pulsed mass spectrometer with time dispersion." Physical Review 69(691).
159. Storey, J. D. and R. Tibshirani (2003). "Statistical significance for genomewide studies." Proc Natl Acad Sci U S A 100(16): 9440-9445.
160. Strittmatter, E. F., L. J. Kangas, K. Petritis, H. M. Mottaz, G. A. Anderson, Y. Shen, J. M. Jacobs, D. G. Camp, 2nd and R. D. Smith (2004). "Application of peptide LC retention time information in a discriminant function for peptide identification by tandem

mass spectrometry." J Proteome Res 3(4): 760-769.

161. Tabb, D. L. (2008). "What's driving false discovery rates?" J Proteome Res 7(1): 45-46.
162. Tabb, D. L., C. G. Fernando and M. C. Chambers (2007). "MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis." J Proteome Res 6(2): 654-661.
163. Tanner, S., H. Shu, A. Frank, L. C. Wang, E. Zandi, M. Mumby, P. A. Pevzner and V. Bafna (2005). "InsPecT: identification of posttranslationally modified peptides from tandem mass spectra." Anal Chem 77(14): 4626-4639.
164. Taylor, C. F. (2006). "Minimum reporting requirements for proteomics: a MIAPE primer." Proteomics 6 Suppl 2: 39-44.
165. Taylor, C. F., P. A. Binz, R. Aebersold, M. Affolter, R. Barkovich, E. W. Deutsch, D. M. Horn, A. Huhmer, M. Kussmann, K. Lilley, M. Macht, M. Mann, D. Muller, T. A. Neubert, J. Nickson, S. D. Patterson, R. Raso, K. Resing, S. L. Seymour, A. Tsugita, I. Xenarios, R. Zeng and R. K. Julian, Jr. (2008). "Guidelines for reporting the use of mass spectrometry in proteomics." Nat Biotechnol 26(8): 860-861.
166. Taylor, C. F., D. Field, S. A. Sansone, J. Aerts, R. Apweiler, M. Ashburner, C. A. Ball, P. A. Binz, M. Bogue, T. Booth, A. Brazma, R. R. Brinkman, A. Michael Clark, E. W. Deutsch, O. Fiehn, J. Fostel, P. Ghazal, F. Gibson, T. Gray, G. Grimes, J. M. Hancock, N. W. Hardy, H. Hermjakob, R. K. Julian, Jr., M. Kane, C. Kettner, C. Kinsinger, E. Kolker, M. Kuiper, N. Le Novere, J. Leebens-Mack, S. E. Lewis, P. Lord, A. M. Mallon, N. Marthandan, H. Masuya, R. McNally, A. Mehrle, N. Morrison, S. Orchard, J. Quackenbush, J. M. Reecy, D. G. Robertson, P. Rocca-Serra, H. Rodriguez, H. Rosenfelder, J. Santoyo-Lopez, R. H. Scheuermann, D. Schober, B. Smith, J. Snape, C. J. Stoeckert, Jr., K. Tipton, P. Sterk, A. Untergasser, J. Vandesompele and S. Wiemann (2008). "Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project." Nat Biotechnol 26(8): 889-896.
167. Taylor, C. F., N. W. Paton, K. S. Lilley, P. A. Binz, R. K. Julian, Jr., A. R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch, M. J. Dunn, A. J. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T. A. Neubert, S. D. Patterson, P. Ping, S. L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T. M. Vondriska, J. P. Whitelegge, M. R. Wilkins, I. Xenarios, J. R. Yates, 3rd and H. Hermjakob (2007). "The minimum information about a proteomics experiment (MIAPE)." Nat Biotechnol 25(8): 887-893.
168. Thomson, J. J. (1910). "Rays of positive electricity." Phil. Mag. Series 20(118): 752-767.
169. Thomson, J. J. (1913). Rays of positive electricity and their application to chemical analysis. London, Longman's Green and Company.
170. Turck, C. W., A. M. Falick, J. A. Kowalak, W. S. Lane, K. S. Lilley, B. S. Phinney, S. T. Weintraub, H. E. Witkowska, N. A. Yates and G. Association of Biomolecula Resource Facilities Proteomics Research (2007). "The Association of Biomolecular Resource Facilities Proteomics Research Group 2006 study: relative protein quantitation." Mol Cell Proteomics 6(8): 1291-1298.
171. Unlu, M., M. E. Morgan and J. S. Minden (1997). "Difference gel electrophoresis: a single gel method for detecting changes in protein extracts." Electrophoresis 18(11): 2071-2077.
172. Vizcaino, J. A., E. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Ríos, J. A. Dianes,

Z. Sun, T. Farrah, N. Bandeira, P. A. Binz, I. Xenarios, M. Eisenacher, G. Mayer, L. Gatto, A. Campos, R. J. Chalkley, H. J. Kraus, J. P. Albar, S. Martinez-Bartolomé, R. Apweiler, G. Omenn, L. Martens, A. R. Jones and H. Hermjakob (2013). "ProteomeXchange: globally co-ordinated proteomics data submission and dissemination (Submitted)." Nat Biotechnol.

173. Vizcaino, J. A., L. Martens, H. Hermjakob, R. K. Julian and N. W. Paton (2007). "The PSI formal document process and its implementation on the PSI website." Proteomics 7(14): 2355-2357.
174. Walzer, M., D. Qi, G. Mayer, J. Uszkoreit, M. Eisenacher, T. Sachsenberg, F. F. Gonzalez-Galarza, J. Fan, C. Bessant, E. W. Deutsch, F. Reisinger, J. A. Vizcaino, J. A. Medina-Aunon, J. P. Albar, O. Kohlbacher and A. R. Jones (2013). "The mzQuantML Data Standard for Mass Spectrometry-based Quantitative Studies in Proteomics." Mol Cell Proteomics 12(8): 2332-2340.
175. Wang, G., W. W. Wu, Z. Zhang, S. Masilamani and R. F. Shen (2009). "Decoy methods for assessing false positives and false discovery rates in shotgun proteomics." Anal Chem 81(1): 146-159.
176. Wang, R., A. Fabregat, D. Rios, D. Ovelleiro, J. M. Foster, R. G. Cote, J. Griss, A. Csordas, Y. Perez-Riverol, F. Reisinger, H. Hermjakob, L. Martens and J. A. Vizcaino (2012). "PRIDE Inspector: a tool to visualize and validate MS proteomics data." Nat Biotechnol 30(2): 135-137.
177. Washburn, M. P., D. Wolters and J. R. Yates, 3rd (2001). "Large-scale analysis of the yeast proteome by multidimensional protein identification technology." Nat Biotechnol 19(3): 242-247.
178. Wasinger, V. C., S. J. Cordwell, A. Cerpa-Poljak, J. X. Yan, A. A. Gooley, M. R. Wilkins, M. W. Duncan, R. Harris, K. L. Williams and I. Humphery-Smith (1995). "Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium." Electrophoresis 16(7): 1090-1094.
179. Wilkins, M. R., R. D. Appel, J. E. Van Eyk, M. C. Chung, A. Gorg, M. Hecker, L. A. Huber, H. Langen, A. J. Link, Y. K. Paik, S. D. Patterson, S. R. Pennington, T. Rabilloud, R. J. Simpson, W. Weiss and M. J. Dunn (2006). "Guidelines for the next 10 years of proteomics." Proteomics 6(1): 4-8.
180. Wilkins, M. R., C. Pasquali, R. D. Appel, K. Ou, O. Golaz, J. C. Sanchez, J. X. Yan, A. A. Gooley, G. Hughes, I. Humphery-Smith, K. L. Williams and D. F. Hochstrasser (1996). "From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis." Biotechnology (N Y) 14(1): 61-65.
181. Wilm, M. S. and M. Mann (1994). "Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last?" International Journal of Mass Spectrometry and Ion Processes 136: 167-180.
182. Wisniewski, J. R., A. Zougman, N. Nagaraj and M. Mann (2009). "Universal sample preparation method for proteome analysis." Nat Methods 6(5): 359-362.
183. Yates, J. R., 3rd, J. K. Eng and A. L. McCormack (1995). "Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases." Anal Chem 67(18): 3202-3210.
184. Yates, J. R., 3rd, J. K. Eng, A. L. McCormack and D. Schieltz (1995). "Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database." Anal Chem 67(8): 1426-1436.

Anexo

ANEXO:

Análisis de los datos del experimento multi-centro 6 de ProteoRed (PME6) con la herramienta MIAPE Extractor.

Como hemos comentado, una de las iniciativas de ProteoRed es organizar diversos experimentos multi-centro en los que un número de participantes, laboratorios españoles de ProteoRed y algunos de fuera de España, analizan una misma muestra con diferentes equipamientos y, en ocasiones, diferentes aproximaciones. Luego, los resultados son recopilados y comparados en una reunión, con la intención de compartir cada una de las aproximaciones con el resto, aprender unos de otros, y sacar las conclusiones oportunas para una mejora de los protocolos de análisis.

En el caso del experimento multi-centro 6 de ProteoRed (PME6), consistía en la caracterización de la muestra ASS17v3 consistente en una mezcla de en torno a 145 proteínas de plasma humano tal y como se describe en la sección de materiales y métodos (sección 3.2.1) junto con 4 proteínas añadidas a diferentes concentraciones (5 órdenes de magnitud de diferencia):

- *P61981|1433_HUMAN* (30 µg),
- *P00883|ALDOA_RABIT* (3 µg),
- *P02666|CASB_BOVIN* (0,3 µg) y
- *P00489|PYGM_RABIT* (0,03 µg).

En total participaron 20 laboratorios, cada uno realizando **tres réplicas técnicas**, para así también analizar la reproducibilidad dentro de cada laboratorio.

Con la herramienta de análisis del MIAPE Extractor se realizaron por tanto dos tipos de análisis:

A. Análisis de los datos enviados por cada participante en las plantillas de resultados aportadas para ello.
B. Análisis centralizado de los datos crudos, utilizando estándares y el motor de búsqueda Mascot.

En la tabla (Tabla 7) se muestran cada uno de los conjuntos de datos que se utilizaron para los dos tipos de análisis (ver columna 'Análisis resultados laboratorios' y 'Análisis centralizado (Mascot), para ver qué conjuntos de datos se utilizaron en un caso y en otro).

	Laboratorio	Espectrómetro	Tipo de espectrómetro	Código de laboratorio	Buscador(es)	FDR en Extractor	FDR reportada	Análisis resultados laboratorios	Análisis centralizado (Mascot)
1	CIMA	Synapt HD	QTOF	QT_1	Phenyx	1.00%	-	x	x
2	CIC bioGUNE	Qtof Micro	QTOF	QT_2	PLGS	-	4.00%	x	
3	UPV	Qtof Premier	QTOF	QT_3	VEMS	1.00%	-	x	
4	CBT	Qstar Elite	QTOF	QT_4	Mascot	1.00%	-	x	
5	CIPF	Qstar XL	QTOF	QT_5	Mascot, Paragon	-	-		x
6	INIBIC	4800 MALDI TOFTOF	MALDI	M_1	Phenyx	-	0.95%	x	x
7	PCM-UCM	4800 MALDI TOFTOF	MALDI	M_2	Paragon	-	2.00%	x	x
8	CNB	4800 MALDI TOFTOF	MALDI	M_3	Mascot, Phenyx	1.00%	-	x	x
9	UA	Ion Trap XCT plus	Ion trap	IT_1	Phenyx, SpectrumMill	-	0%, 1%	x	x
10	VH	Ion trap Esquire-Ultra	Ion trap	IT_2	Mascot	-	0.00%	x	x
11	CNB	HCT Ultra 3D	Ion trap	IT_3	Mascot	1.00%	-	x	x
12	PCM-UCM	LTQ	Linear Ion Trap	LIT_1	Mascot	-	0.00%	x	x
13	LP-CSIC UAB	LTQ-Orbi	Orbi	O_1	Sequest, OMSSA, Phenyx	-	0.3%, 0.8%, 0.8%	x	x
14	VH	LTQ-Orbi	Orbi	O_2	Mascot	1.00%	-	x	x
15	CMU	LTQ-Orbi-Velos	Orbi Velos	OV_1	Phenyx	-	1.00%	x	x
16	UPF	LTQ-Orbi-Velos	Orbi Velos	OV_2	Mascot	1.00%	-	x	x
17	PCB	LTQ-Orbi-Velos	Orbi Velos	OV_3	PeptideProphet	-	1.30%	x	x
18	UB	LTQ-Orbi-Velos	Orbi Velos	OV_4	Sequest	-	4.70%	x	x
19	CIB	LTQ-Orbi-Velos	Orbi Velos	OV_5	Mascot, Sequest	-	-		x
20	UCO	LTQ-Orbi-Velos	Orbi Velos	OV_6	Sequest	-	-		x

Tabla 7. Datos de participantes en el PME6 seleccionados para el análisis: De un total de 20 conjuntos de datos, 17 fueron seleccionados para la comparativa de los resultados enviados por los laboratorios ('x' en columna 'Análisis resultados laboratorios') y otros 17 fueron seleccionados para la comparativa del re-análisis centralizado de los datos ('x' en columna 'Análisis centralizado'). Cada conjunto de datos tiene asociado un código de laboratorio indicando el tipo de espectrómetro utilizado (QT: QTOF, IT: Ion Trap, M: MALDI, LIT: Linear Ion Trap, O: Orbitrap, OV: Orbitrap Velos). También se muestran los buscadores que utilizaron los laboratorios en sus análisis.

Todas las gráficas mostradas en este documento están directamente exportadas de la herramienta MIAPE Extractor. El principal objetivo de las comparativas aquí descritas es mostrar la utilidad de la herramienta para este tipo de análisis, por lo que cada gráfica se describirá manera concisa.

A. Análisis de los datos enviados por cada participante

De los 20 participantes, se tomaron las hojas Excel de resultados disponibles en http://www.proteored.org/PME6_Results.asp y se introdujeron en el MIAPE Extractor como se describió anteriormente (sección 4.2.5), incorporando la información proteína-péptido-puntuación de péptido. En este proceso, del cual se incorporaron datos de hasta 9 motores de búsqueda diferentes, se encontraron los siguientes problemas con los que tuvimos que trabajar, adaptando, en su caso, la herramienta para ello:

- Pese a que los participantes usaban una misma plantilla, existían aún diferencias en el formato de los datos que iban a imposibilitar la comparación.
    - Algunos participantes obtuvieron las proteínas con identificadores (ID) de Uniprot, en vez de códigos de acceso (AC) (ej. *CASB_BOVIN* en vez de *P02666*). Para solucionarlo, la herramienta traduce automáticamente los códigos ID de Uniprot a los códigos AC correspondientes, únicos para cada isoforma.
    - Cada motor de búsqueda exporta con un formato diferente las secuencias peptídicas: incluyendo las modificaciones en la secuencia, incluyendo los aminoácidos anterior y posterior separados por "-" o por ".", etc… Se adaptó la herramienta para traducir los diferentes formatos de secuencias peptídicas a un único formato consistente únicamente en la secuencia peptídica tríptica.
    - Algunos motores de búsqueda exportan en sus resultados grupos de proteínas, separadas por comas, a los que pertenece un mismo péptido. Para tener en cuenta dichos grupos también se adaptó la herramienta para considerarlas proteínas distintas, ya que el agrupamiento se haría posteriormente con el algoritmo PAnalyzer (Prieto, Aloria et al. 2012).
- Algunos participantes no incluyeron los PSMs identificados en los resultados, si no únicamente los péptidos, es decir, la mejor asignación de cada secuencia peptídica (el mejor PSM para cada péptido). Esto, aunque no supone ningún problema, debemos tenerlo en cuenta para no comparar los datos a nivel de PSM.
- Algunos participantes no incluyeron en los resultados los péptidos junto con las proteínas a las que pertenecían. Para solucionarlo se implementó un sistema por el cual

cuando la herramienta detecta que no existe tal relación, pide al usuario seleccionar el fichero FASTA con el cual se hicieron las búsquedas, para así buscar los códigos de acceso de las proteínas y las secuencias peptídicas, para construir automáticamente las relaciones, con todas sus ambigüedades (péptidos pertenecientes a varias proteínas) si las hubiera.

- Cada tipo de puntuación (Mascot score, XCorr de SEQUEST, e-value de OMSSA) fue codificada en el fichero de texto que se extrajo de cada resultado para tenerla en cuenta a la hora de ordenar las listas de péptidos para calcular las tasas de error (FDR).
- Además, no hubo un criterio común para enviar los resultados con una tasa de error FDR concreta, por lo que cada participante optó por una tasa de error diferente en muchos casos.
- De hecho además, algunos participantes no incluyeron las identificaciones DECOY en las plantillas, por lo que en su caso no fue posible hacer un corte por FDR controlado con la herramienta. Sin embargo, las variaciones de la tasa de error van desde un 1% a un 5%, que en el caso del análisis de una muestra con supuestamente unas 145 proteínas, representan un rango de entre 1 hasta 5 proteínas erróneas, lo cual no debería influir en la comparativa.
- Pese a que todos los participantes debían usar una misma base de datos creada para el experimento, sospechamos que algunos de ellos, utilizando herramientas/buscadores no muy comunes, no usaron dicha base de datos.

Pese a intentar incorporar la mayoría de los resultados obtenidos por los 20 participantes, finalmente se optó por no incluir en la comparativa algunos de ellos, debido a disparidades de números junto con la imposibilidad de normalizarlos por un criterio de corte por FDR, lo cual se achaca a posibles interpretaciones erróneas de la plantilla de resultados, o posibles problemas en los propios análisis de las muestras (Tabla 7). Finalmente se seleccionaron un total de 17 conjuntos de datos, los cuales se clasificaron con un código según el tipo de espectrómetro de masas utilizado (Figura 56).

Los datos de esta comparativa contendrán una variabilidad inherente a cada espectrómetro de masas, además de la variabilidad con respecto a los análisis bioinformáticos realizados por cada participante. En la parte B de este anexo, compararemos los datos de los participantes re-analizados con Mascot y por tanto, eliminando la variabilidad derivada de los análisis bioinformáticos.

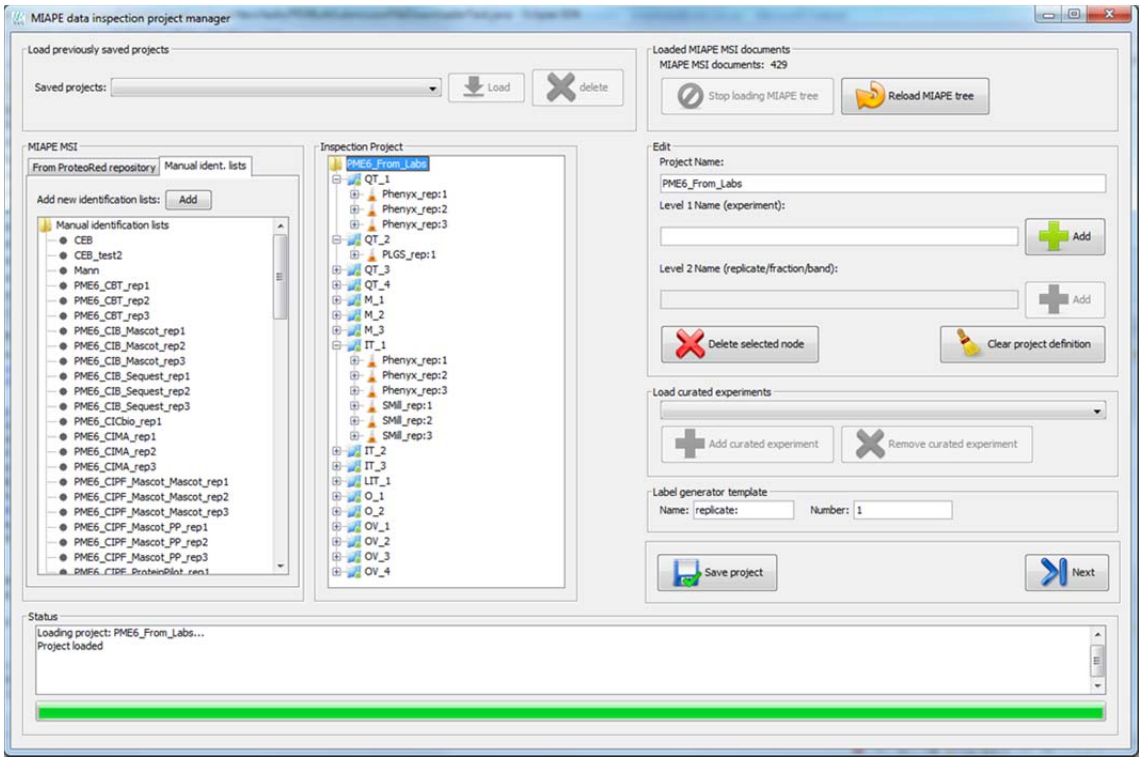


Figura 56. Proyecto de inspección de los resultados obtenidos por los participantes del PME6: Tras incorporar las tablas Excel de resultados en la herramienta, se creó un proyecto de inspección llamado "PME6_From_Labs" con los datos que se pudieron incorporar de 17 de los 20 laboratorios. Algunos de ellos, como el IT_1, integrando los resultados de distintos motores de búsqueda, en este caso, Phenyx y Spectrum Mill.

Filtrado de datos

Los datos se cargaron en el MIAPE Extractor, obteniéndose la siguiente Figura 57 (A) representando el número de péptidos y proteínas. Ciertos datos se pudieron filtrar por 1% de FDR a nivel de péptido (B) gracias a que los participantes incluyeron las identificaciones señuelo (con códigos de acceso con el prefijo "*rev_*"). Los que no, se dejaron tal cual, asumiendo las tasas de error enviadas por cada participante (Tabla 7), salvo el caso del QT_2, al que se le hizo un corte de puntuación >=5 (del valor de la puntuación del *ProteinLynx Global Server*) para intentar bajar un poco la tasa de error de un 4% con la que los datos habían sido enviados. El corte se hizo en 5 para filtrar los datos fuera de la distribución de puntuaciones como se muestra en la Figura 58.

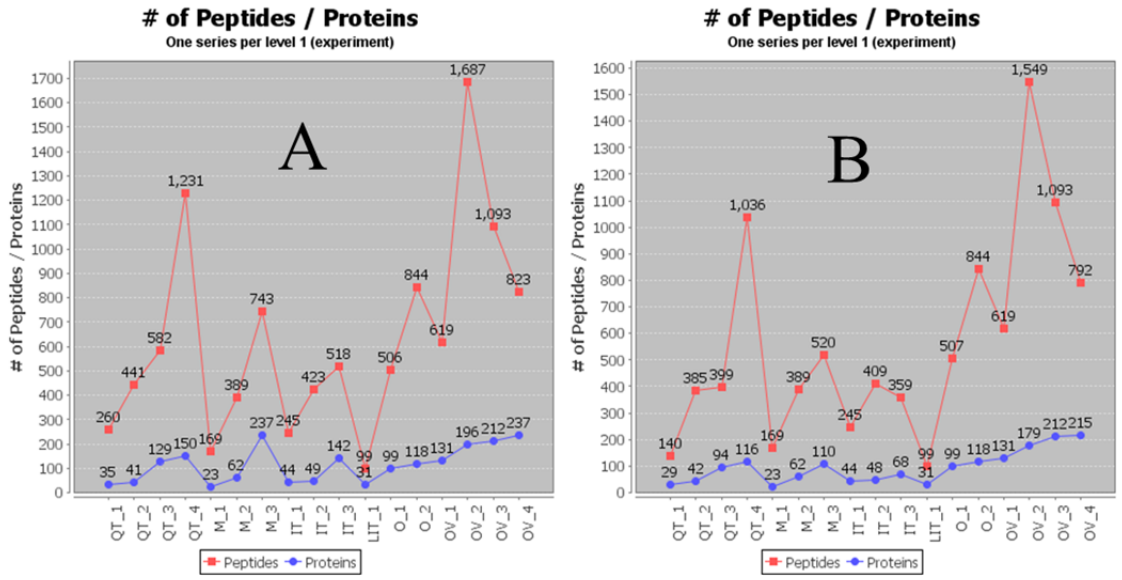


Figura 57. Número de proteínas péptidos: Número de proteínas (línea azul) y de péptidos (línea roja) (secuencias peptídicas diferentes) identificados por cada participante, sin filtrar (A) y filtrando según describe en el texto (B).

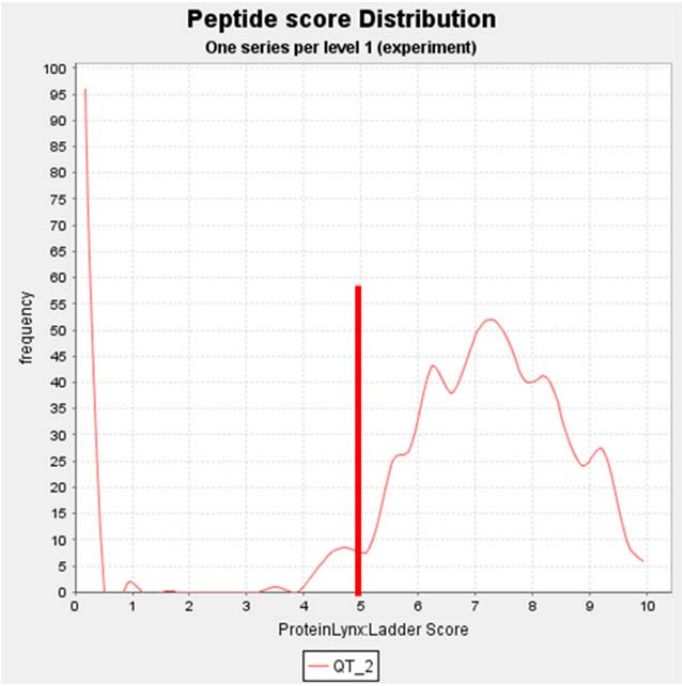


Figura 58. Distribución de la puntuación del QT_2: Se filtró por un valor de la puntuación para péptidos del buscador *ProteinLynx Global Server* mayor o igual que 5 para descartar las puntuaciones fuera de la distribución.

Número de identificaciones y variabilidad entre réplicas

En la Figura 57 B, podemos ver el número de identificaciones que obtuvo cada participante, tanto en número de péptidos como en número de proteínas. Sin embargo, dado que cada participante realizó 3 réplicas, y algunos de ellos analizaron cada una de las réplicas con varios buscadores, es interesante también ver la reproducibilidad de sus análisis viendo la media y la desviación estándar de dichos números (Figura 59). El experimento QT_2 no muestra variabilidad ya que únicamente empleó una réplica.

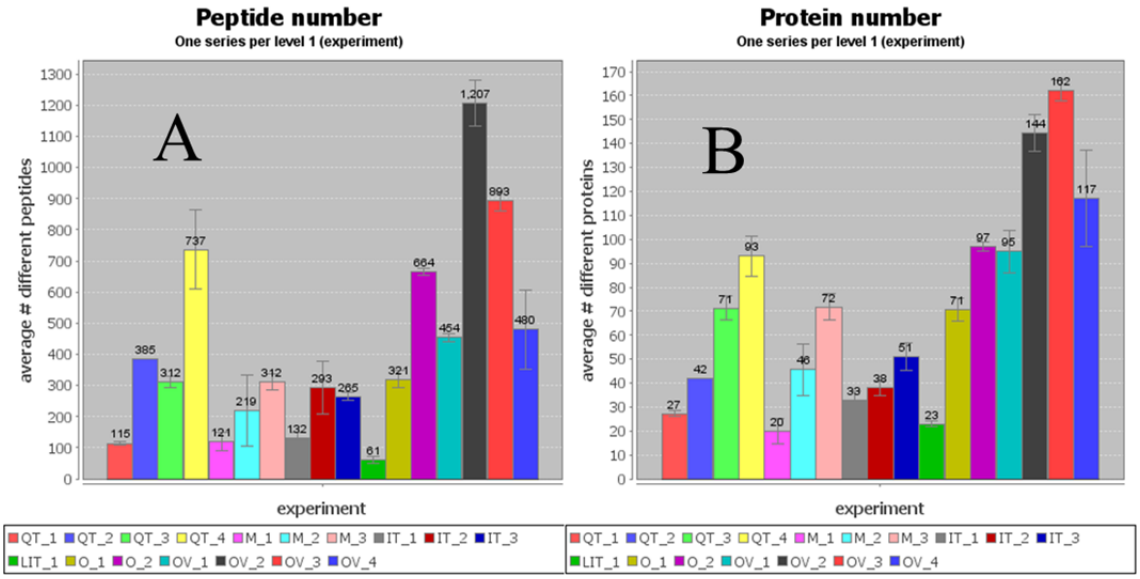


Figura 59. Número medio y desviación estándar de identificaciones de péptidos y proteínas: Para cada participante se muestra el número medio de péptidos (A) y proteínas (B) obtenidos sobre las tres réplicas o sobre los diferentes análisis sobre las 3 réplicas.

Destacan los resultados el laboratorio O_1 (LTQ Orbitrap), que pese a incluir los resultados provenientes de 9 conjuntos de datos (al utilizar 3 buscadores distintos sobre las 3 réplicas), su variabilidad es mínima (más en detalle en la Figura 60), incluso menor de lo que cabría esperar en los resultados de buscadores diferentes sobre un mismo conjunto de espectros. Esto se explica por la aplicación de la herramienta de desarrollo propio *Integrator* para integrar los resultados de varios motores de búsqueda (http://proteomica.uab.cat/index.php?option=com_content&view=article&id=80). Por otra parte, vemos en la Figura 59 que pese a que el laboratorio OV_2 obtiene un número de péptidos mayor, el experimento OV_3 tiene un rendimiento mayor en cuanto a número de proteínas, habiendo utilizado los dos un Orbitrap Velos y el buscador Mascot. El mayor rendimiento en OV_3 se podría explicar por la utilización del conjunto de herramientas *Trans Proteomics Pipeline* (Keller y Shteynberg 2011), tras las búsquedas en Mascot, con los algoritmos de

Peptide Prophet y *Protein Prophet*, que sobre todo con éste último, se mejora sensiblemente el rendimiento de la inferencia de proteínas.

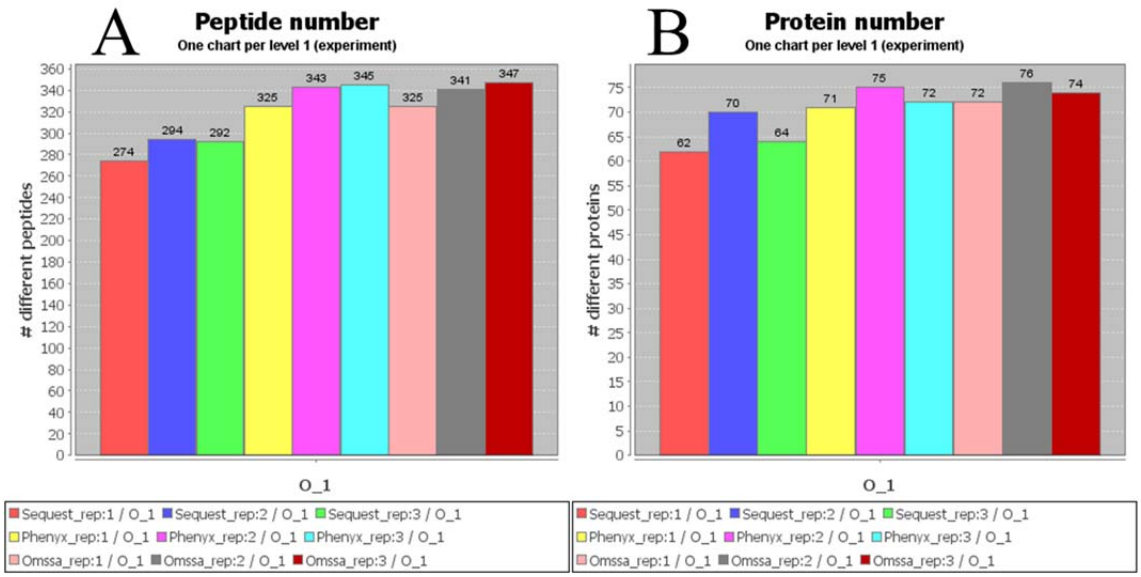


Figura 60. Número de péptidos (A) y proteínas (B) de participante O_1: Número de péptidos y proteínas del participante O_1 (LTQ Orbitrap), para cada una de las 3 réplicas analizadas con los motores de búsqueda SEQUEST (rojo claro, azul oscuro y verde), Phenyx (amarillo, rosa y azul claro) y OMSSA (rosa claro, gris y rojo).

También podemos concluir el alto rendimiento general en número de identificaciones junto con un alto grado de reproducibilidad de los Orbitraps (LTQ Orbitraps y Orbitraps Velos) en comparación con otros espectrómetros, únicamente comparable en este caso a los datos aportados por QT_3 y QT_4 (QTOFs) y M_3 (MALDI).

Comparación de puntuaciones

Otra de las comparaciones que se pueden hacer en el MIAPE Extractor consiste en la superposición de las distribuciones de puntuaciones obtenidas por cada motor de búsqueda. En el caso de Mascot podemos ver en la Figura 61 (A) cómo la distribución de OV_2 destaca sobre las demás. Para normalizar la comparación el usuario puede seleccionar "RELATIVE_FRECUENCY" como tipo de histograma, para así obtener la gráfica de la Figura 61 (B). Como se puede observar (gráfica aumentada en B), las mejores puntuaciones corresponden con M_3 (azul oscuro) y LIT_1 (rosa oscuro), siendo las puntuaciones de OV_2 de peor calidad (rosa claro).

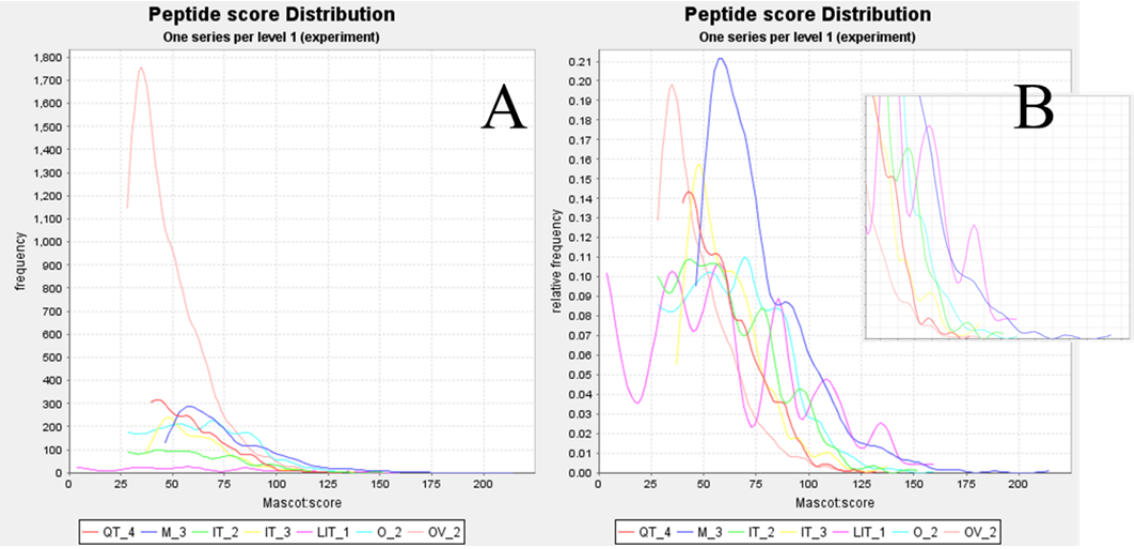


Figura 61. Distribuciones de puntuaciones Mascot (Mascot:Score): Distribuciones de las puntuaciones obtenidas por los participantes que usaron el motor de búsqueda Mascot (A): QT_4 (rojo), M_3 (azul oscuro), IT_2 (verde), IT_3 (amarillo), LIT_1 (rosa), O_2 (azul claro) y OV_2 (rosa claro). Normalización relativa de los histogramas (B).

En el caso del motor de búsqueda Phenyx, representando la puntuación *Z-score*, obtenemos la Figura 62 (A). Tras normalizar de la misma manera (B) vemos claramente cómo QT_1 (rojo) tiene puntuaciones de mayor calidad mientras que OV_1 (azul claro) de peor calidad que otros (pese a tener bastante mayor número de ellos, como se ve en A y en la Figura 57.

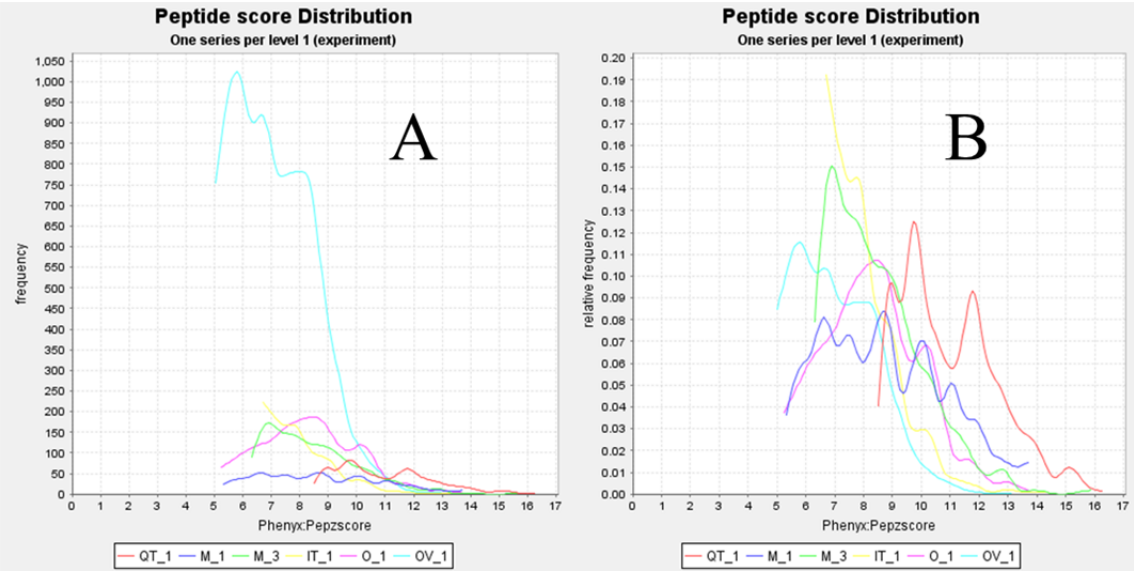


Figura 62. Distribuciones de puntuaciones de Phenyx (Z-score): Distribuciones de las puntuaciones obtenidas por los participantes que usaron el motor de búsqueda Phenyx (A): QT_1 (rojo), M_1 (azul oscuro), M_3 (verde), IT_1 (amarillo), O_1 (rosa) y OV_1 (azul claro). Normalización relativa de los histogramas (B).

En el caso del buscador SEQUEST (Figura 63), únicamente tenemos dos participantes, O_1 y OV_4 y parece que OV_4 (azul) obtiene unas puntuaciones de calidad ligeramente mayor que O_1 (rojo). En este caso, la normalización no realizaba cambios significativos, ya que las dos distribuciones son ya comparables entre sí.

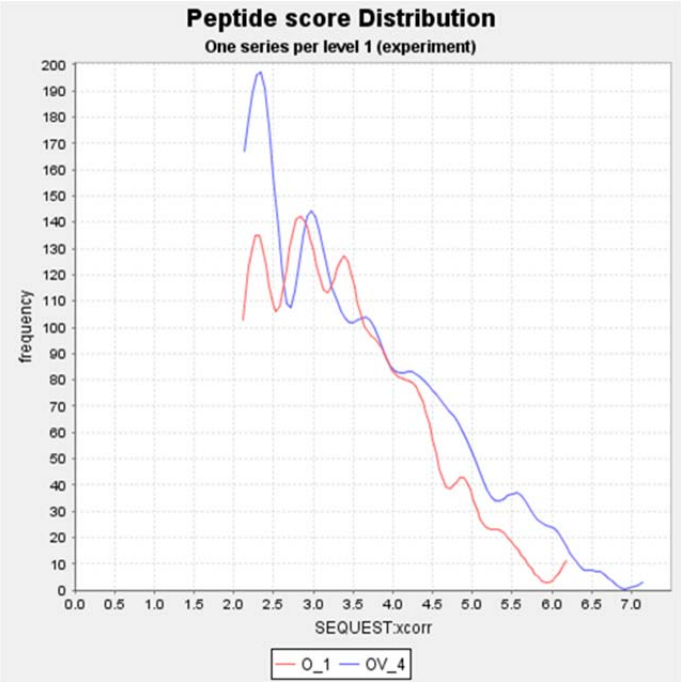


Figura 63. Distribuciones de puntuaciones Sequest (XCorr): Distribuciones de las puntuaciones obtenidas por los participantes que usaron el motor de búsqueda Sequest. En rojo, las puntuaciones de O_1 y en azul las de OV_4. En este caso no fue necesario normalizar las frecuencias ya que los dos histogramas son claramente comparables.

La herramienta MIAPE Extractor permite además comparar las puntuaciones de tal manera que se muestra para cada péptido en común de entre dos conjuntos de datos, cuál es la mejor puntuación que se ha obtenido. Es precisamente lo que vemos en la Figura 64. En A, vemos la comparación de las puntuaciones XCorr obtenidas por cada péptido en O_1 y en OV_4. En este caso, la tendencia global es que para una misma secuencia peptídica, la mejor puntuación sea más alta en O_1 que en OV_4, en contraposición con la gráfica anterior. Hay que tener en cuenta que lo mostrado en la Figura 64 sólo son los péptidos en común y el solapamiento de péptidos es de sólo el 27.2% (Figura 65). En el caso de la Figura 64 B, se muestran las comparaciones dos a dos de las tres réplicas del experimento M_3, lo que muestra que las puntuaciones son bastante reproducibles, ajustándose mejor a la recta diagonal (negra).

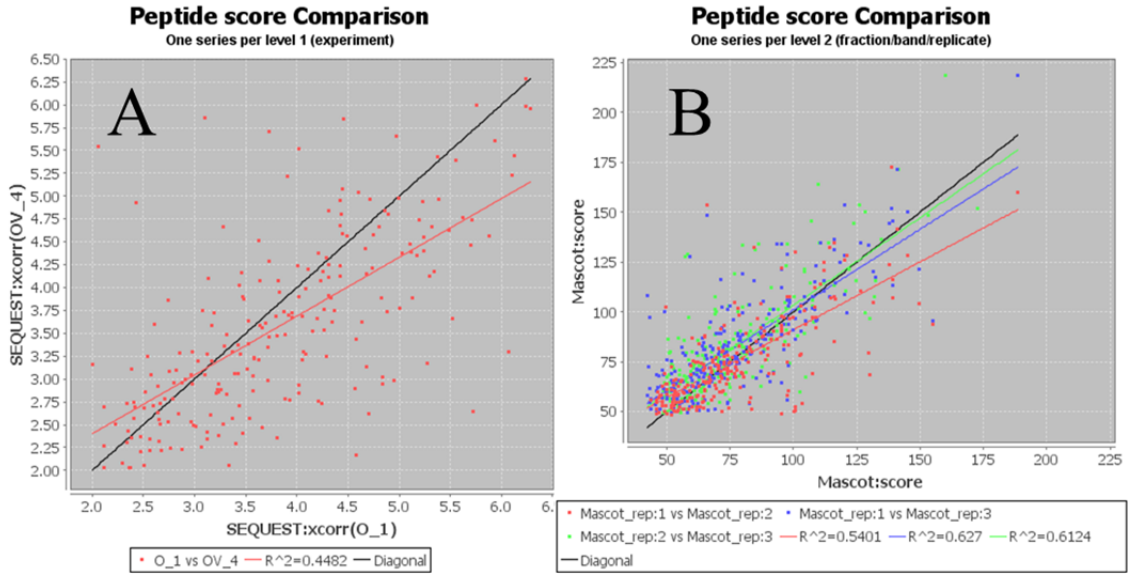


Figura 64. Comparación valores de XCorr (Sequest) entre O_1 y OV_4 y comparación valores Mascot:score entre réplicas de M_3: (A) Comparación entre las puntuaciones XCorr obtenidas por los mismos péptidos en O_1 (eje de abscisas) y OV_4 (eje de ordenadas). (B) Comparación entre las puntuaciones Mascot:score obtenidas por los mismos péptidos en las tres réplicas de M_3. La línea negra muestra la diagonal. Las línea rojas, azul y verde muestran la tendencia global en cada comparación, y en la leyenda se muestra el error cuadrático del ajuste a dichas rectas.

Solapamientos

El MIAPE Extractor, también es capaz de mostrar cuál es el solapamiento entre 2 o 3 conjuntos de datos tanto de péptidos como de proteínas, creando diagramas de Venn (Figura 65). En el caso del solapamiento entre O_1 (naranja) y OV_4 (verde), podemos ver en la Figura 65 que el solapamiento a nivel de péptidos es de un 27.2% y un 19.8% a nivel de proteína. Además también muestra que sólo el 22.4% de los péptidos en O_1 son exclusivos, es decir, que no están en OV_4, mientras que un en OV_4 los péptidos exclusivos son un 50.3%. De la misma manera, sólo el 17.9% de las proteínas de O_1 no están en OV_4, mientras que el 62.2% de las proteínas en OV_4 son exclusivas comparándolas con O_1. Es además clara la diferencia en el número de identificaciones, como se observa en la diferencia de tamaño de las circunferencias del diagrama de Venn (un 77.6% de la unión de los dos conjuntos pertenece a OV_4 mientras que sólo un 49.7% pertenece a O_1).

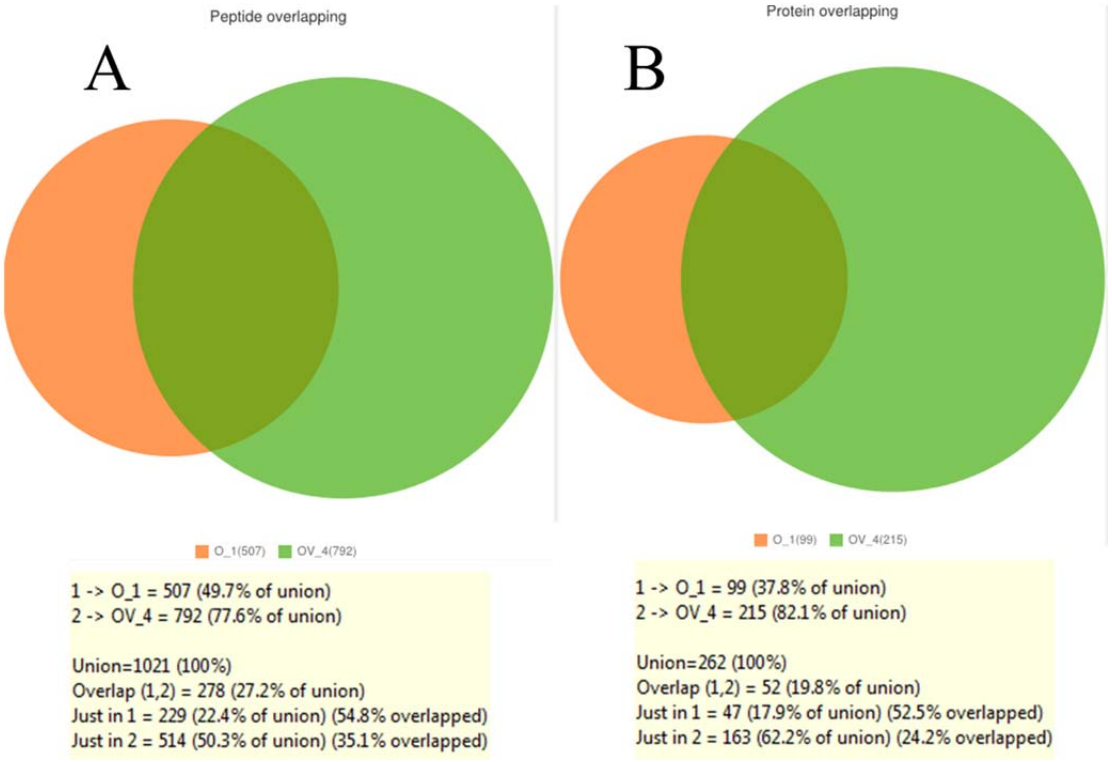


Figura 65. Solapamiento entre péptidos (A) y proteínas (B) entre O_1 y OV_4: El solapamiento a nicel de péptido es del 27.2% y a nivel de proteína del 19.8%.

De la misma manera podemos ver el solapamiento entre las réplicas de un mismo experimento, como es el caso del solapamiento a nivel de proteína de las 3 réplicas del experimento OV_3 y entre las réplicas del experimento OV_4 (Figura 66). Como se puede observar en la Figura 66 (A), para OV_3, el solapamiento de dos en dos réplicas es 63-75%, y de un 56.6% para las tres réplicas. Sin embargo, en la Figura 66 (B), para OV_4, vemos que el solapamiento es claramente menor, siendo de 32-39% si comparamos las réplicas de dos en dos y de un 22.7% para las tres réplicas. Estos resultados concuerdan con la alta desviación típica de OV_4 de la Figura 59.

Figura 66. Solapamiento de proteínas entre las réplicas de OV_3 (A) y entre las réplicas de OV_4 (B): Se muestra el solapamiento en términos de número de proteínas entre las 3 réplicas del experimento OV_3 (A) y las réplicas del experimento OV_4 (B).

Reproducibilidad inter-laboratorio

Para analizar la reproducibilidad conjunta de los laboratorios, es útil la representación del *heat map*, una matriz representada por una escala de colores que indica el número de veces que se ha detectado el péptido o la proteína representada en dicha fila. En la Figura 67 se muestran los péptidos (A y B) y proteínas (C), desde los más vistos arriba hasta los menos vistos abajo. Cada columna representa los datos de cada laboratorio, en el mismo orden que gráficas anteriores.

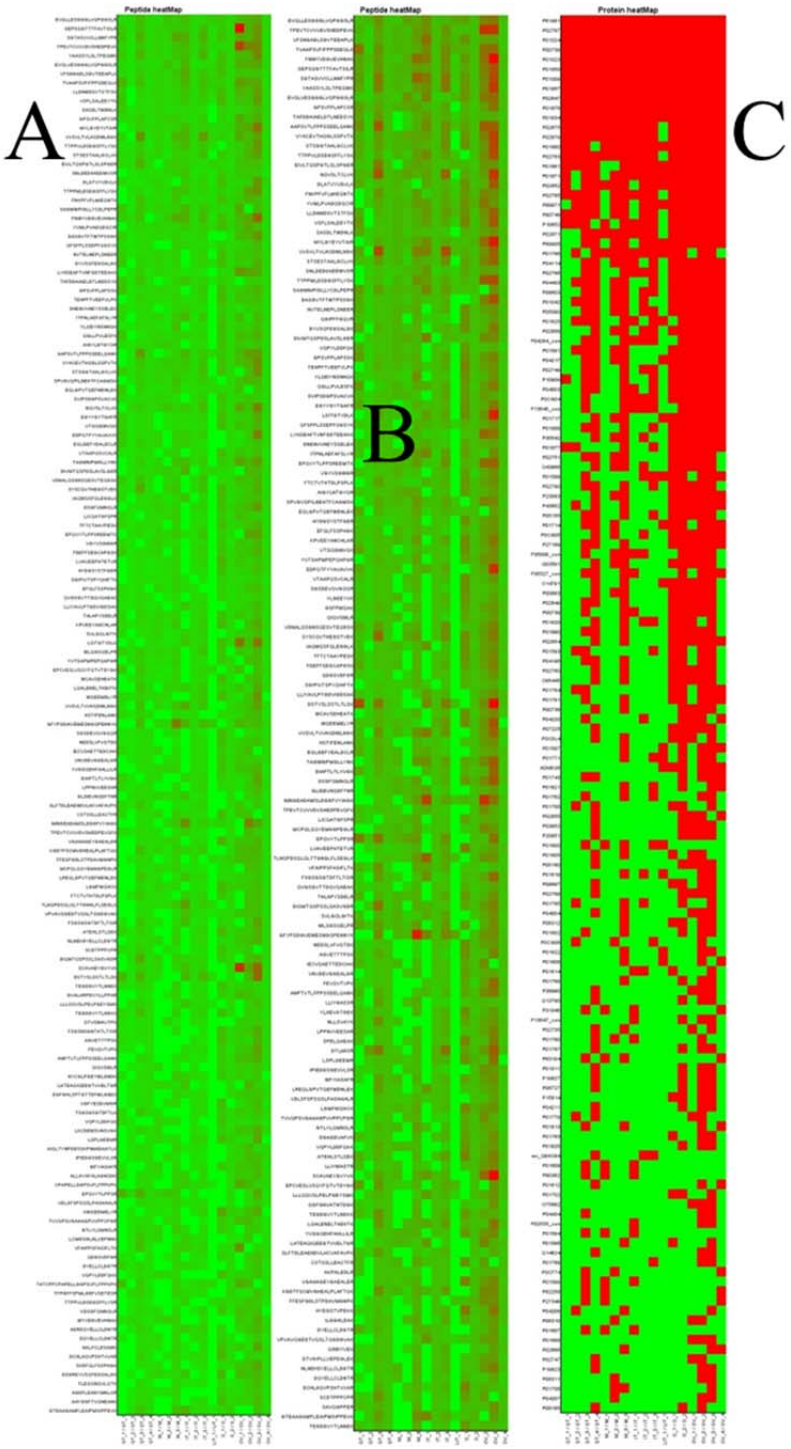


Figura 67. Heat maps de péptidos (A y B) y proteínas (C): La escala de colores representa el número de veces que se ha visto un péptido o una proteína (en orden ascendente de verde a rojo). En el caso del heat map de péptidos se muestran los péptidos que se han visto hasta en 12 de experimentos, la escala de colores es logarítmica en los dos casos A y B. En el caso de A, existe un péptido que es visto 1640 veces por el experimento OV_1. Dicho péptido enmascara el color del resto, incluso en escala logarítmica. En B se muestran los mismos datos, excluyendo los datos de OV_1, por lo que se puede apreciar más detalle en la reproducibilidad de los péptidos. En C, se muestran las proteínas que se han visto hasta en 3 experimentos.

Otra manera de ver las diferencias entre las identificaciones de los laboratorios es ver, en este caso, las proteínas que sólo han sido vistas en cada uno de los experimentos comparados. En la Figura 68 vemos precisamente eso. En A se puede observar que los Orbitrap Velos (OV_x) tienen el mayor número de proteínas exclusivas. En B, se muestra también la tendencia acumulativa de número de proteínas, según vamos haciendo la unión de las proteínas de los sucesivos experimentos. La herramienta también permite ver el mismo tipo de gráfica pero con respecto a los péptidos exclusivos de cada experimento.

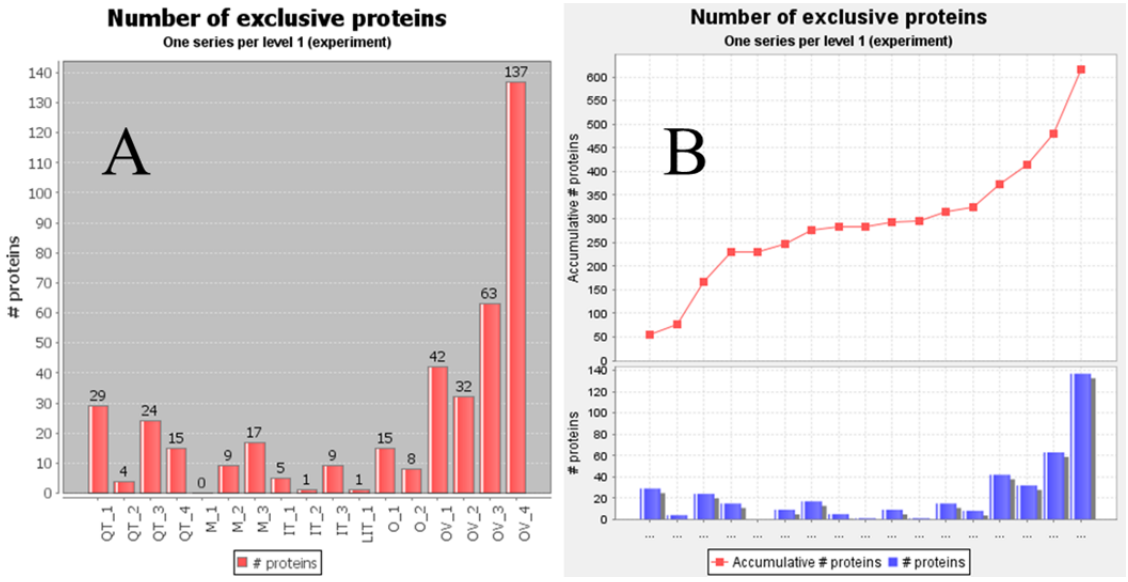


Figura 68. Proteínas exclusivas de cada experimento: En A se muestran el número de proteínas que sólo se han visto en cada uno de los experimentos. En B, se muestra una gráfica compuesta por el mismo diagrama de barras abajo y arriba se muestra la tendencia acumulativa de ir haciendo la unión de las proteínas de manera sucesiva en el orden mostrado.

Rendimiento de la digestión

Para evaluar el rendimiento de la digestión, tenemos la representación de los números de péptidos con ciertos puntos de corte ausentes (*missed-cleavages*) por la tripsina. En la Figura 69 vemos el número de péptidos detectados en cada experimento con ningún *miss-cleavage* (rojo), uno (azul oscuro), dos (verde), tres (amarillo), cuatro (rosa) y cinco o más (azul claro). Se observa un rendimiento menor de las digestiones en los casos de los experimentos M_2, LIT_1, OV_2 y OV_4, con cerca de un 50% de los péptidos con puntos de corte omitidos.

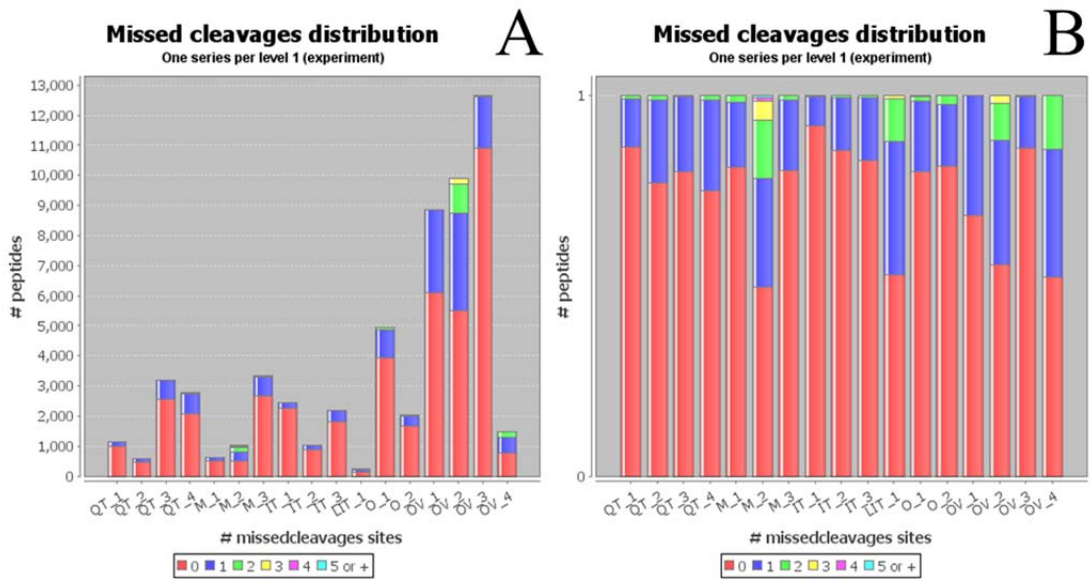


Figura 69. Comparación de puntos de corte ausentes (*missed-cleavages*): En A y B se muestran el número de péptidos diferentes que tienen 0 (rojo), 1 (azul oscuro), 2 (verde), 3 (amarillo), 4 (rosa) o 5 o más (azul claro) puntos de corte ausentes (*missed-cleavages*) de tripsina. En A se muestran los números absolutos. En B se muestran los valores normalizados a 1.

Tamaño de los péptidos

Otro aspecto a inspeccionar con la herramienta es el tamaño de los péptidos. Esto se puede hacer tanto por la masa teórica calculada de las secuencias peptídicas como por la distribución de longitudes. En la Figura 70 (A) vemos la distribución de masas de los experimentos QT_4, M_3, IT_3, LIT_1, O_2 y OV_4 (por simplificar la gráfica, se seleccionó un experimento para cada tipo de espectrómetro). Podemos ver cómo el experimento LIT_1 es el que tiene el mayor número de péptidos con más longitud (el 53% de sus péptidos tienen 20 o más aminoácidos) (Figura 70 B), lo cual concuerda con los datos de la Figura 69 que indican un bajo rendimiento de la digestión. Por otro lado el experimento OV_4 (Figura 70 A) detecta los péptidos más largos, dentro del rango de 4.000 y 4.800 Dalton.

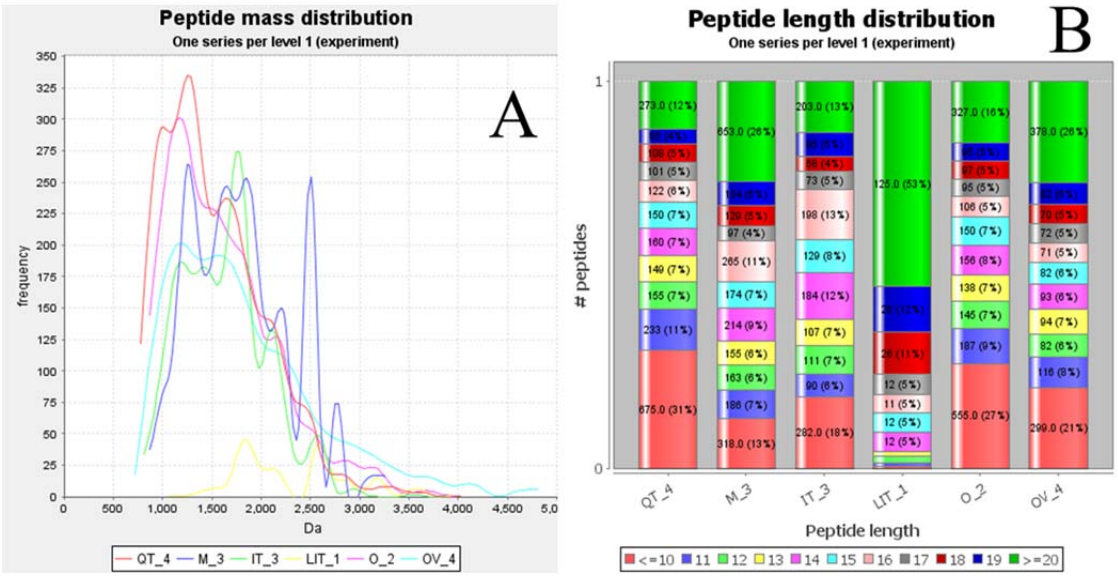


Figura 70. Distribución de masas y de longitudes de péptidos: En A vemos la distribución de masas en Dalton de los 6 experimentos seleccionados. En B, vemos para cada experimento, el número de péptidos con longitud menor o igual que 10, con longitud 11, 12...19 o mayor o igual que 20. Los límites de longitud máxima y mínima son configurables en la herramienta MIAPE Extractor.

Péptidos por proteína

Otro aspecto a inspeccionar con la herramienta es cuántas proteínas se han visto con un solo péptido, cuántas con 2, etc… Esto lo podemos ver en la Figura 71, donde se muestran el número de proteínas identificadas con tan sólo un péptido (rojo), con dos (azul), con tres (verde) o con cuatro o más (amarillo), refiriéndose a secuencias peptídicas diferentes. En nuestro ejemplo podemos destacar el alto porcentaje de proteínas identificadas con un péptido en el experimento OV_4 (lo que también explicaría su variabilidad en sus réplicas), en contraposición con los experimentos QT_2 y M_1, donde pese a que el número de proteínas identificadas es bastante menor (A), el porcentaje de proteínas identificadas con más de un péptido es muy alto.

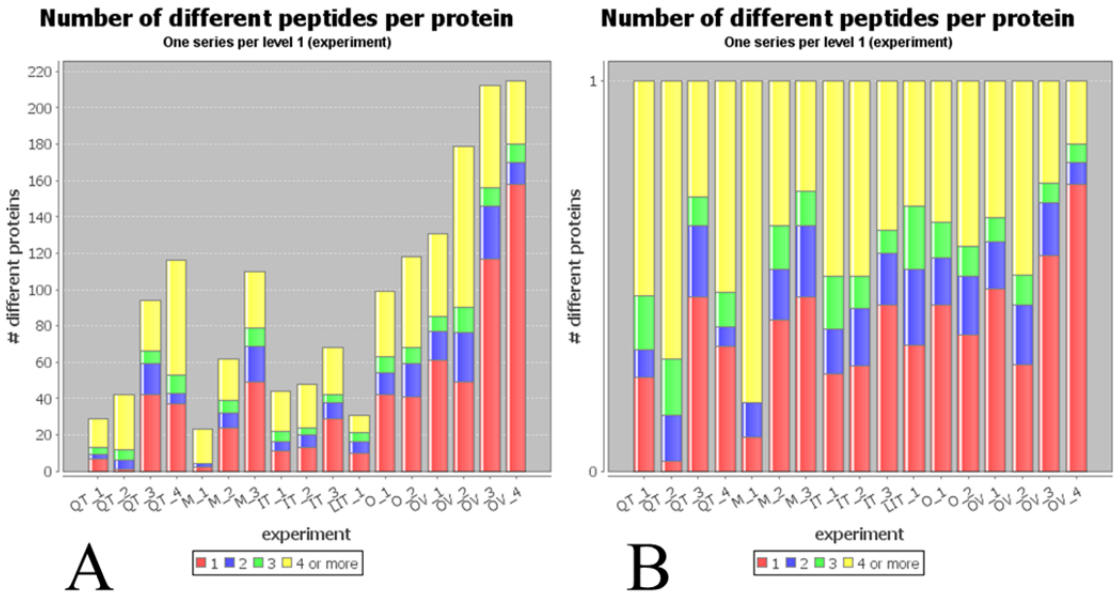


Figura 71. Número de proteínas identificadas en función del número de péptidos asociados a ellas: En A vemos el número absoluto de proteínas identificadas en cada experimento con 1 sólo péptido diferente (rojo), con 2 péptidos diferentes (azul), con 3 péptidos diferentes (verde) o con 4 o más péptidos diferentes (amarillo). En B vemos representados los mismos datos, normalizando los valores a 1.

Redundancia en identificaciones

En numerosas ocasiones una misma secuencia peptídica o una misma proteína es identificada en varias ocasiones, o en varios experimentos. Esto lo podemos ver en la Figura 72, donde en A y B vemos cuántos péptidos se han visto 1 (rojo), 2 (azul), 3 (verde) o 4 o más veces (amarillo) en cada experimento a lo largo de todas sus réplicas. Podemos destacar aquí al experimento O_1, en el que se ve que todos los péptidos han sido vistos al menos dos veces, y la mayoría de hecho, se ha visto más de 4 veces. Recordemos que este laboratorio utilizó la herramienta *Integrator* para combinar los resultados de búsqueda de Mascot, Phenyx y OMSSA, por lo que queda claro que uno de los criterios utilizados en dicho software fue no incluir péptidos que no se hubieran visto en al menos dos búsquedas. Un rendimiento también alto muestra el experimento M_1, en el que un porcentaje muy pequeño de los péptidos sólo se han visto una vez. Todo lo contrario que el experimento QT_2 y OV_4, en el que un alto porcentaje de los péptidos sólo se han visto una vez. Por otro lado, si no queremos saber cuántas veces se ha visto un péptido, si no que queremos ver en cuántas réplicas se ha visto, tenemos la representación de la Figura 72 C y D. Ahí podemos ver cómo O_1 destaca por tener una gran reproducibilidad. M_2 y, de nuevo, OV_4 tienen un rendimiento menor que el resto, excluyendo a QT_2 ya que todos sus péptidos se han visto sólo en una réplica, ya que este laboratorio sólo adquirió una vez la muestra y lo analizó con un único buscador.

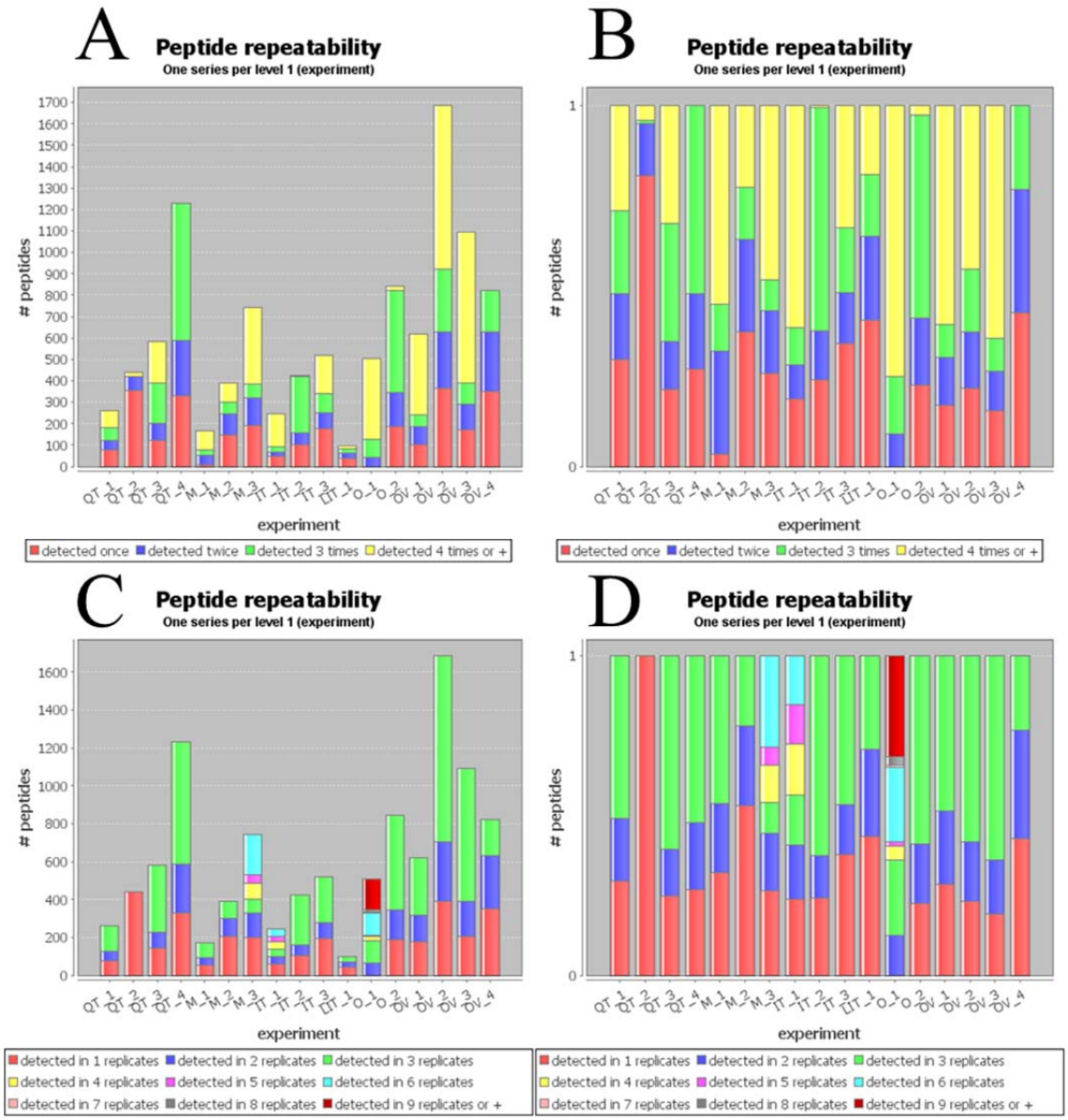


Figura 72. Repetitividad de péptidos: En A y B se muestran para cada experimento el número de péptidos que se han visto una vez (rojo), dos veces (azul), tres veces (verde) o más de tres veces (amarillo) a lo largo de todas sus réplicas. En A se muestran los valores absolutos y en B se muestran los valores normalizados a 1. En C y D se muestran para cada experimento el número de, en este caso, réplicas en las que los péptidos han sido vistos. En C se muestran los valores absolutos y en D se muestran los valores normalizados a 1.

Cobertura de secuencia de proteínas

Uno de los aspectos que suele interesar al investigador es la cobertura de secuencia que tienen sus proteínas. La cobertura se puede calcular, teniendo la secuencia de la proteína y superponiendo las secuencias de sus péptidos para ver qué porcentaje de la secuencia proteica está cubierta por los péptidos. En el caso de la herramienta MIAPE extractor, en principio la

información de la secuencia de la proteína no está disponible, ya que no es una información requerida por las directrices MIAPE MSI. Sin embargo, siempre que se trabaje con códigos de acceso de Uniprot, el usuario tendrá la posibilidad de dejar que la herramienta busque las secuencias en Internet, y calcule las coberturas automáticamente. Así pues, podemos ver en la Figura 73 (A), la media y desviación típica de la cobertura de secuencia de las proteínas identificadas, y en (B), las distribuciones de coberturas de secuencia de cada experimento.

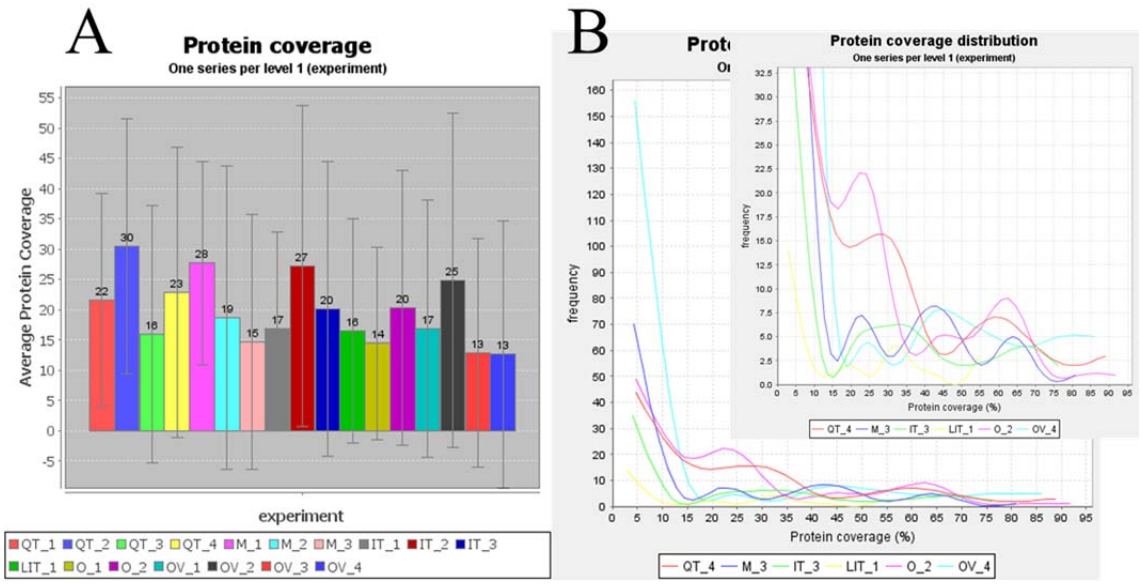


Figura 73. Cobertura de secuencia de las proteínas de cada experimento: En A podemos ver la cobertura de secuencia media (y la desviación típica) de todas las proteínas identificadas en cada experimento. En B, vemos la distribución de coberturas de 6 experimentos seleccionados (por simplificar).

Correspondencia con los cromosomas humanos

Otra utilidad de la herramienta consiste en saber de forma inmediata el número de genes que han codificado a las proteínas inspeccionadas en el proyecto. En la Figura 74 (A) se muestra para 6 experimentos seleccionados (uno para cada tipo de espectrómetro), la cobertura de los genes detectados (por una proteína a la que codifica) de cada cromosoma, es decir, el porcentaje de los genes del cromosoma que han codificado proteínas detectadas en la muestra. En este caso, normalizado al experimento que mayor cobertura tuviese de uno de los cromosomas y omitiendo la cobertura del cromosoma Y, ya que al ser muy alta, enmascaraba el resto de valores. Se puede observar que el experimento OV_4 (azul claro en A) tiene una mayor cobertura que el resto en la mayoría de los cromosomas. Por otro lado, en (B), vemos para el conjunto de datos totales (los 6 experimentos seleccionados), cuántas proteínas vienen de cada uno de los genes. En este caso, el 14% de las proteínas provienen de genes pertenecientes al cromosoma 1, seguidos de l9% pertenecientes al cromosoma 19. En (C) vemos el número de

PSMs (azul) y péptidos (rojo) que se han detectado pertenecientes a proteínas de cada cromosoma. Por su parte, la gráfica (D), fue expresamente creada para su utilización en la primera fase del proyecto del consorcio español para el análisis de las proteínas codificantes del cromosoma 16 humano (Sp-HPP), en el que un subconjunto de las proteínas del cromosoma 16 se asignaron a cada grupo.

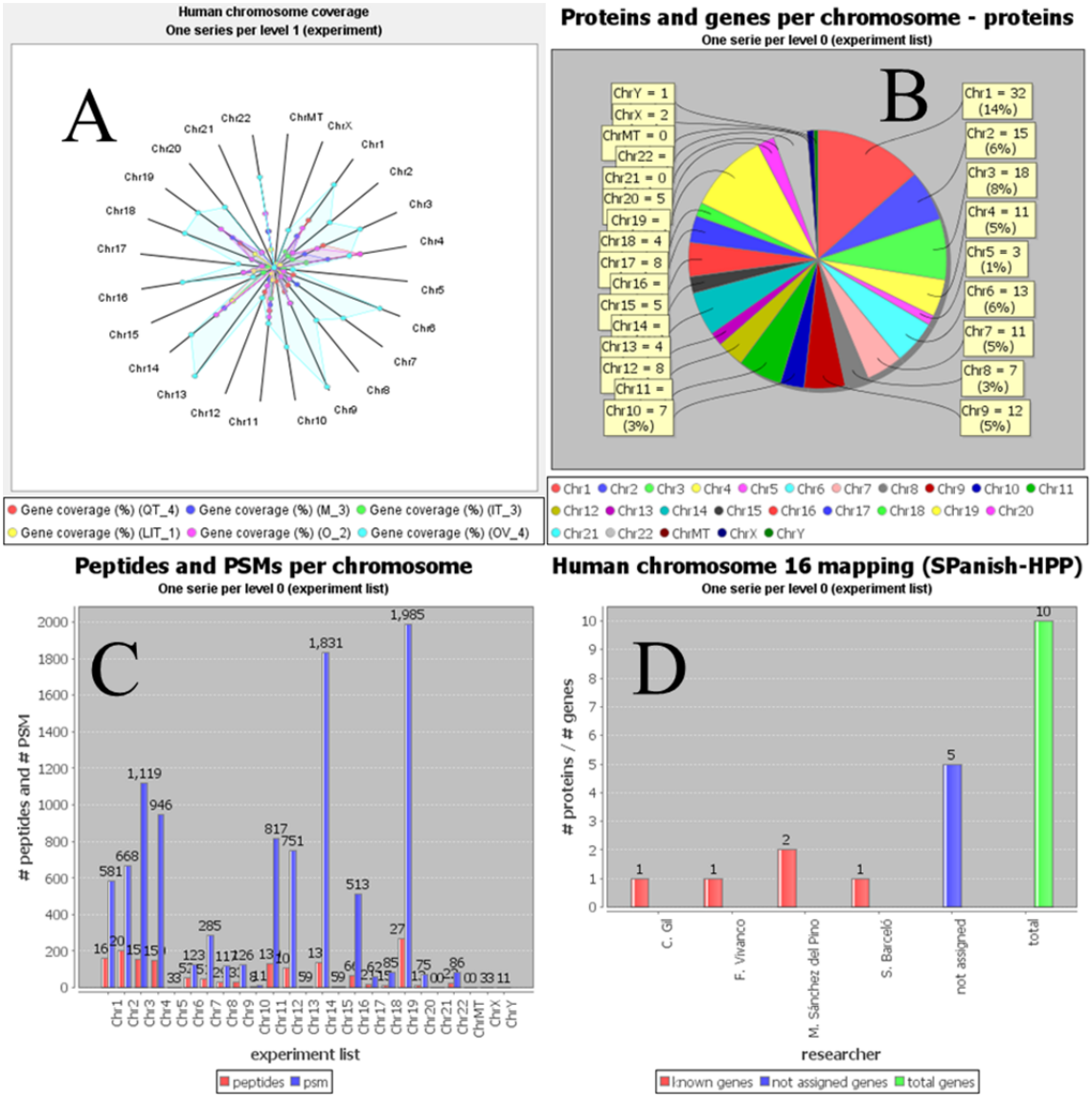


Figura 74. Correspondencia con cromosomas humanos: Estas gráficas muestran de diferentes maneras el número de proteínas que han sido codificadas por los genes de cada uno de los cromosomas humanos. En A se muestran la cobertura del cromosoma cubierta por cada experimento (normalizada por el que tiene mayor cobertura). En B se muestra el porcentaje de proteínas codificadas por cada uno de los cromosomas en el conjunto de datos total. En C se muestran el número de péptidos y PSMs correspondientes a proteínas pertenecientes a cada cromosoma humano. Por último, la gráfica D muestra los genes del cromosoma 16 que han sido asignados a cada grupo del consorcio español del cromosoma humano (Sp-HPP), y los que no.

Sensibilidad para detectar las proteínas "spiked"

Como hemos comentado, a la muestra consistente en una mezcla de complejidad media de proteínas de plasma humano se les añadió 4 proteínas a diferentes concentraciones para evaluar la sensibilidad de cada aproximación. Aplicando un filtro por código de acceso de proteínas con los 4 códigos de acceso: P61981 (30 μg), P00883 (3 μg), P02666 (0,3 μg) y P00489 (0,03 μg), podemos centrarnos exclusivamente en dichas proteínas.

En la Figura 75 se muestran en las gráficas superiores el número de proteínas identificadas de las 4 en cuestión. En los heat-maps inferiores se muestran las proteínas detectadas por cada laboratorio. Se puede observar en (A) cómo la proteína más abundante, la P61981 (Proteína 14-3-3 gamma) es detectada por todos los participantes. La siguiente proteína en abundancia, la P00883 (Aldolasa fructosa bifosfato A) es detectada por 9 de los 17 participantes. Luego, la proteína P02666 (Beta caseína) únicamente es detectada por 3 laboratorios (M_3, OV_3 y OV_4). Sin embargo, en el caso de la proteína presente en menor concentración, la P00489 (Glicógeno fosforilasa) no se detecta en ningún caso. Con el fin de determinar si el corte por FDR ha sido demasiado restringente, quitamos los filtros por FDR y score descritos anteriormente, asumiendo los cortes proporcionados por cada participante según su criterio. Los resultados obtenidos se muestran en la Figura 75 (B), y vemos que la única diferencia es la detección de la proteína P00883 en el experimento IT_3. Inspeccionando esa proteína en ese experimento, vemos que dicha proteína se identificó únicamente con 1 péptidos con una puntuación de Mascot de 31.06, estando cerca del corte de FDR 1% a nivel de péptido.

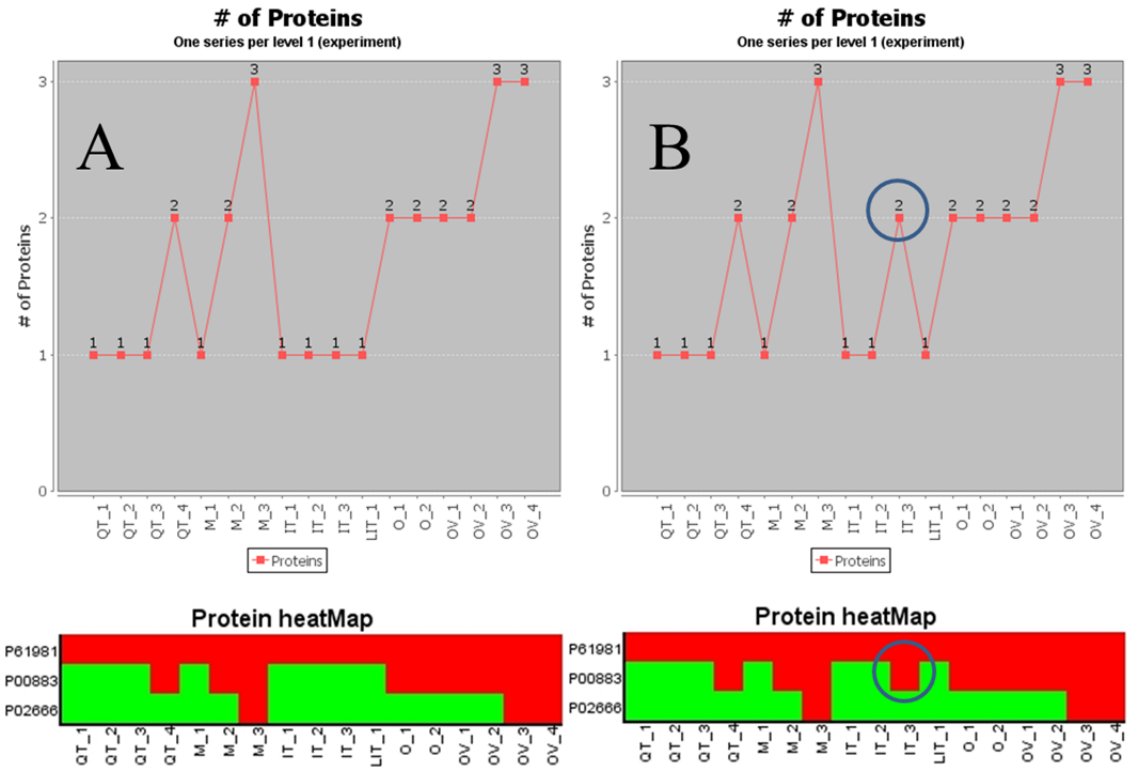


Figura 75. Análisis de las 4 proteínas añadidas a la muestra a distintas concentraciones: Se muestran el número de proteínas detectadas por cada participante (arriba), y el heat-map para las 3 proteínas detectadas (abajo). En A se muestran los datos tras aplicar los filtros descritos anteriormente y en B, sin aplicar ningún filtro.

Sensibilidad y precisión

Existen varios estimadores del rendimiento de las identificaciones, y en general de los test estadísticos. Por ejemplo, la sensibilidad se define como la capacidad de un estimador de clasificar como positivas las identificaciones realmente verdaderas. La precisión indica el porcentaje de identificaciones realmente positivas dentro de las clasificadas como positivas. Se calculan de la siguiente manera:

$$Sesibilidad = \frac{VP}{\text{VP+FN}} \quad Precisión = \frac{TP}{\text{TP+FP}}$$

siendo VP="Verdaderos positivos", FN="Falsos Negativos" y FP="Falsos positivos". La sensibilidad es por tanto la fracción de verdaderos positivos y la precisión el número de identificaciones correctas entre todas las clasificadas como correctas.

Obviamente, para saber qué identificaciones son las verdaderamente correctas y cuáles no, es necesario saber la composición real de la muestra. En este caso, dada la complejidad de la muestra no es posible saberlo con una precisión del 100%. Sin embargo, una aproximación

válida puede considerar coger las proteínas identificadas en más de un laboratorio como proteínas reales en la muestra. Así pues, se cogieron los códigos de acceso de las 141 proteínas identificadas en al menos dos laboratorios y se introdujeron en la herramienta como los verdaderos positivos, para saber los valores de sensibilidad y precisión (Figura 76). Podemos ver que la sensibilidad, es decir, la fracción de verdaderos positivos identificados de los 141 existentes, es alta en la mayoría de los casos (92-100%). Quizás este valor esté sobreestimado debido a la aproximación de coger las proteínas vistas en al menos dos experimentos. En cuanto a la precisión, es decir, el número de identificaciones verdaderamente correctas dentro de las identificaciones que pasan el corte, los valores no son tan extremos, salvo en el caso de M_1 siendo el 100% de sus proteínas verdaderas positivas. Destaca en lo negativo OV_4, con un 32% de verdaderos positivos en sus proteínas consideradas como positivas, acorde con la dispersión vista en varias gráficas anteriores. La mayoría de los participantes sin embargo tienen una precisión oscilando entre 55% y 80%.

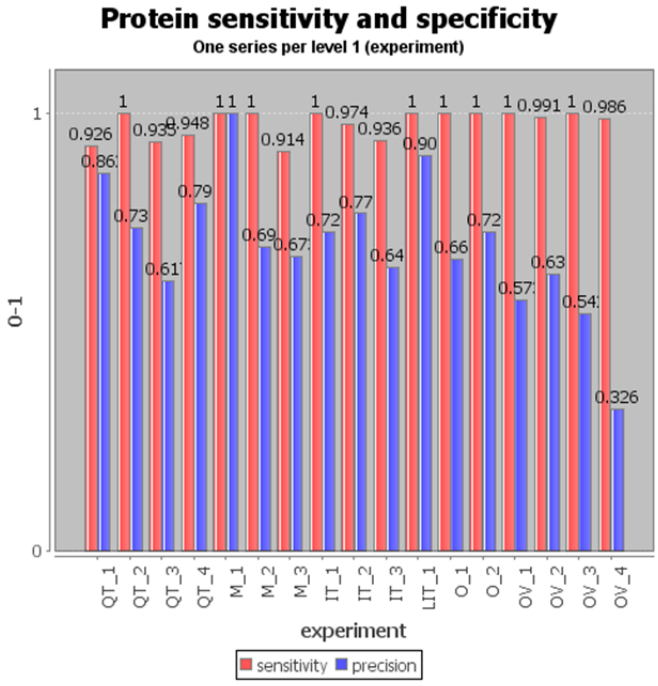


Figura 76. Sensibilidad y precisión de los resultados de los participantes del PME6. Considerando como verdaderas positivas las proteínas que se vieron en al menos dos laboratorios, mostramos la sensibilidad (rojo) y la precisión (azul) de cada uno de los participantes.

B. Análisis centralizado de los datos

Para el análisis centralizado de los datos, se cogieron los datos crudos almacenados en el servidor de ProteoRed y se convirtieron al estándar mzML (Figura 77-1) como se comenta en la sección de materiales y métodos (sección 3.2.1). En el caso de la conversión de algunos de ellos se tuvieron algunos problemas (ya sea porque el mzML resultante era demasiado grande como para utilizarlo como entrada en el buscador Mascot, y en esos casos se intentó conseguir un fichero de listas de picos MGF (*Mascot Generic Format*). Aun así, hubieron 3 conjuntos de datos que no se pudieron procesar, o bien porque el mzML y el MGF no se pudo obtener, o bien porque el mzML obtenido era demasiado grande para utilizarlo luego como entrada en Mascot. Así pues, finalmente se consiguió re-analizar 17 de los 20 conjuntos de datos.

Los ficheros mzML o MGF de los 17 laboratorios se utilizaron como entrada en el buscador Mascot (Figura 77-2), utilizando los parámetros de búsqueda que utilizó cada participante en su propio análisis. Los resultados de todos las búsquedas se exportaron luego al estándar mzIdentML (Figura 77-3).

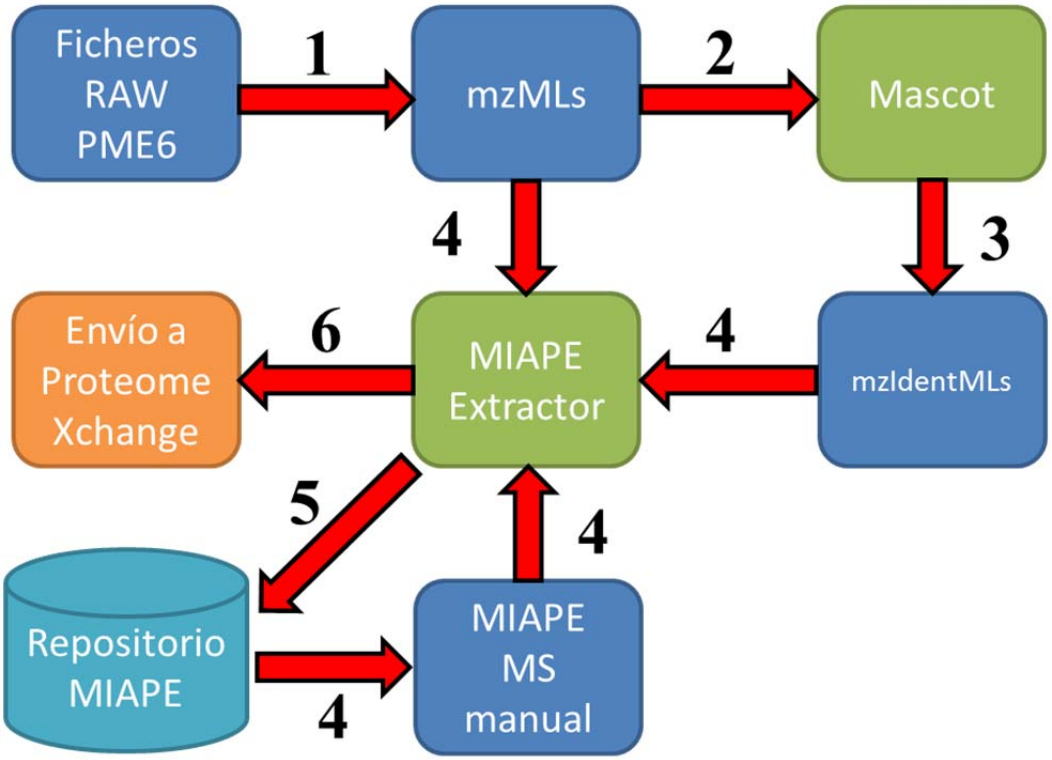


Figura 77. Flujo de re-análisis de los datos del experimento PME6: Los ficheros binarios RAW procedentes de los espectrómetros de masas se convirtieron al estándar mzML (1). Los ficheros mzML resultantes se utilizaron como entrada para su análisis en el motor de búsqueda Mascot (2). Los resultados obtenidos de cada una de las búsquedas se exportaron al estándar mzIdentML (3). Los ficheros mzML, los ficheros mzIdentML y los ficheros MIAPE MS XML (utilizados como plantilla de metadatos MS) fueron utilizados en el MIAPE Extractor (4) para crear los documentos MIAPE en el repositorio (5). Por último, tras el filtrado y análisis de los datos, éstos fueron enviados al repositorio público ProteomeXchange (6), creándose todos los ficheros necesarios automáticamente.

Creación de documentos MIAPE MS y MSI

Usando la opción de la extracción MIAPE en "*batch*" (en lote), se crearon un total de 49 documentos MIAPE MS y 49 documentos MSI, en aproximadamente 3 o 4 horas, de manera desatendida, utilizando como entrada los ficheros mzML, los ficheros mzIdentML y los documentos MIAPE MS XML exportados del repositorio del proyecto PME6 que los participantes crearon manualmente (Figura 77-4). Dichos documentos se almacenaron en el repositorio de documentos MIAPE de ProteoRed en el proyecto con identificador 1191 (Figura 77-5). Las plantillas de metadatos MS utilizadas para crear los documentos fueron los propios documentos MIAPE MS creados manualmente por los participantes en los cuales se incluían las rutas a los ficheros RAW binarios de cada réplica de análisis por espectrometría de masas. Estos ficheros RAW se incluyeron, como se describe más adelante, en el envío de datos de este experimento al ProteomeXchange (Figura 77-6).

Una vez creados, se creó un proyecto de inspección de datos (Figura 78) incluyendo las tres réplicas en cada uno de los experimentos (rep1, rep2 y rep3), excepto QT_5, con una sola réplica, con la codificación propuesta y se inició la inspección de los datos. En la Figura 79 (A) vemos claramente cómo los Orbitrap obtienen un número mayor de PSMs.

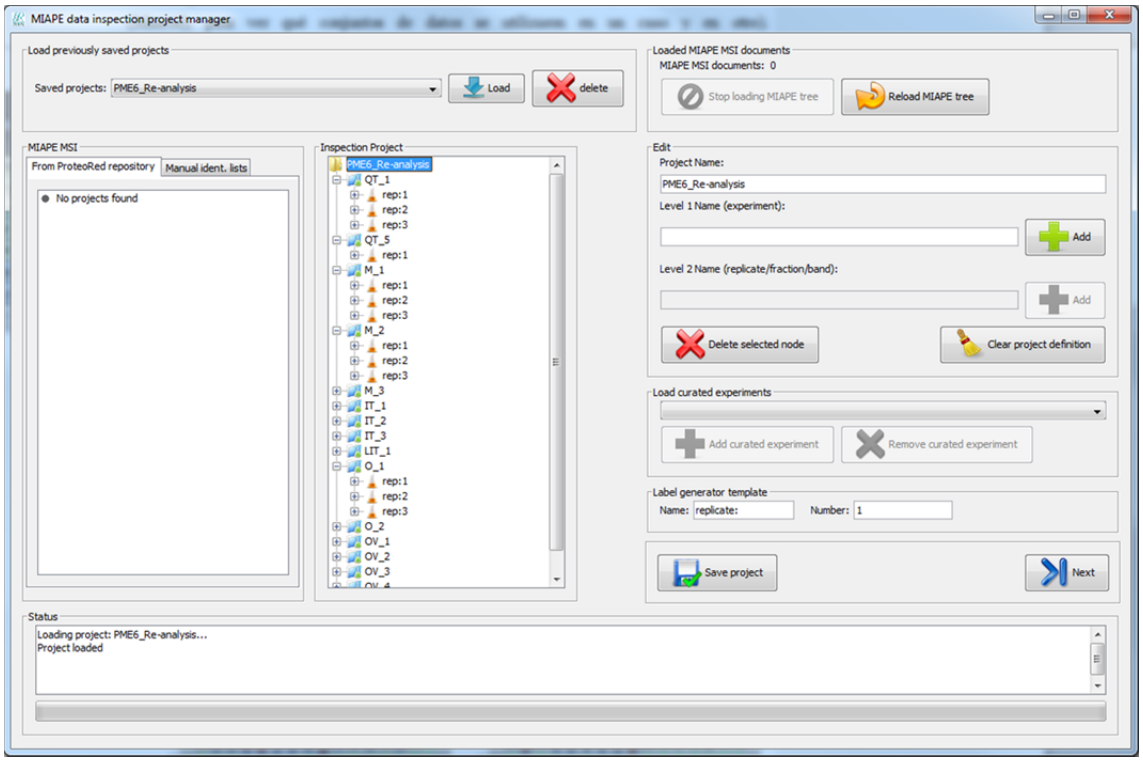


Figura 78. Proyecto de inspección de los resultados de re-análisis de los datos del PME6: Tras crear los documentos MIAPE, se diseñó un proyecto de inspección llamado "PME6_Re-analysis" con los datos de 17 de los 20 laboratorios. Todos ellos con 3 réplicas (rep1, rep2 y rep3), excepto QT_5 que sólo adquirió una réplica (rep1).

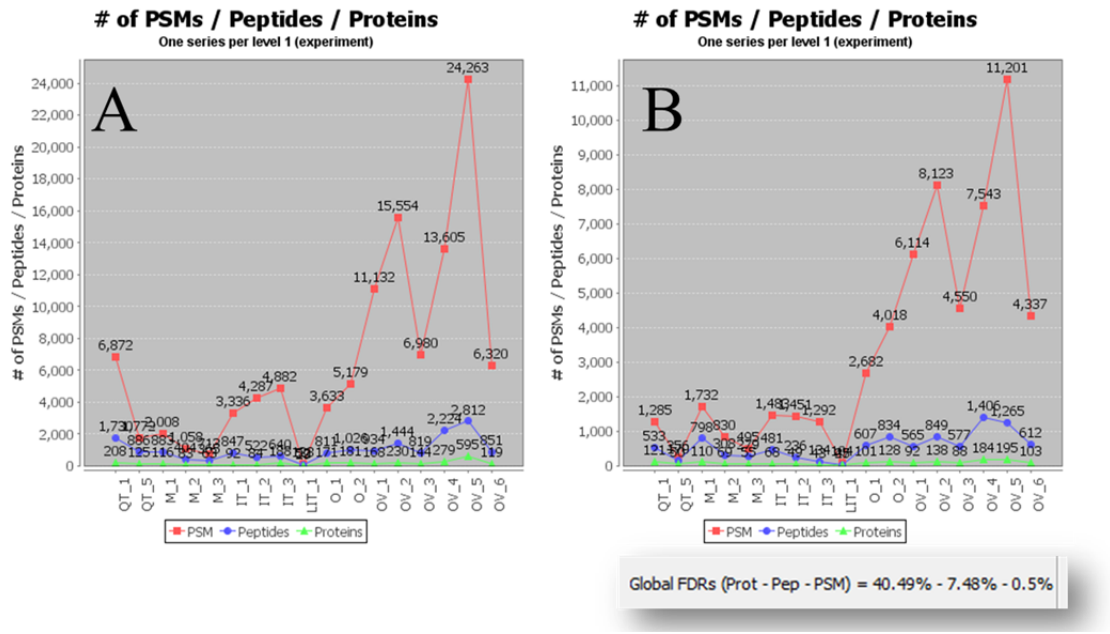


Figura 79. Número de proteínas, péptidos y PSMs: Número de proteínas, péptidos y PSMs identificados por cada experimento, sin filtrar (A) y filtrando por FDR 1% a nivel de péptido (B). Además, gracias a que todos los conjuntos de datos fueron buscados con el motor de búsqueda Mascot, es posible calcular las FDRs globales, mostrada debajo de B.

Filtrado de datos

En este caso, y a diferencia de los datos presentados por cada participante en las plantillas de resultados, se filtraron todos por una tasa de error (FDR) de un 1% a nivel de péptido (Figura 80) y se obtuvieron los resultados mostrados en la Figura 79 (B), con números más o menos proporcionales a los mostrados en (A) sin filtrar.

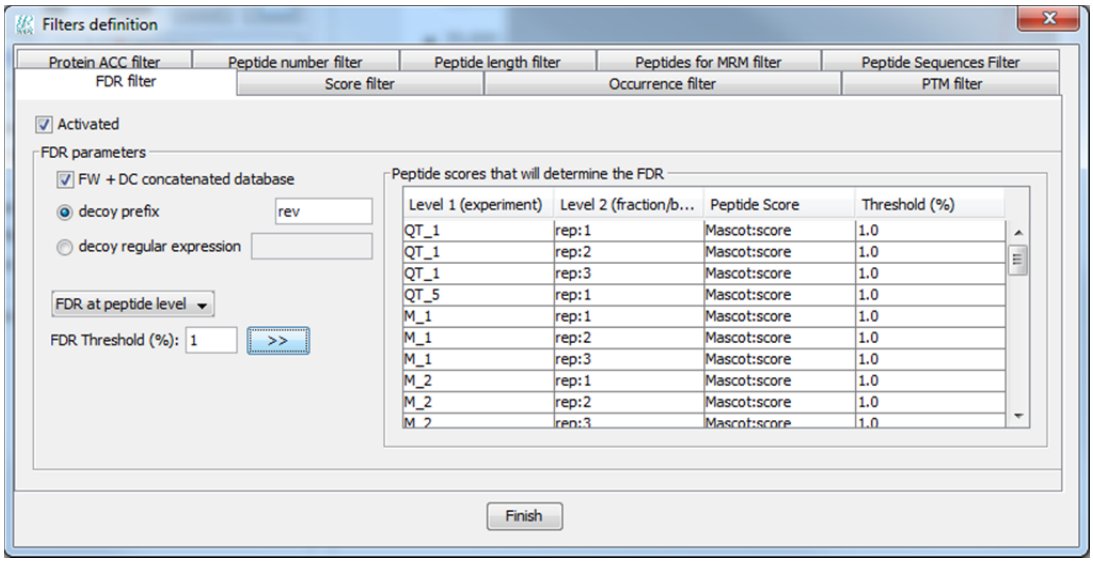


Figura 80. Interfaz gráfica para la aplicación del filtro por FDR. En este caso, la puntuación "Mascot:score" fue seleccionada para aplicar la FDR a nivel de péptido y con un 1% de valor de corte.

Gracias a que todos los datos re-analizados provenían del buscador Mascot, las FDRs globales pudieron ser calculadas por la herramienta, cogiendo las proteínas y péptidos resultantes de la agregación de las tres réplicas de cada participante, haciendo de nuevo el ranking global y calculando la tasa de error resultante: 39.39% a nivel de proteína, 7.28% a nivel de péptido y 0.5% a nivel de PSM. Al aplicar un corte por FDR a nivel de péptido a cada una de las réplicas de los experimentos, quizás se esperase obtener un valor de tasa de error en torno al 1% a nivel de péptido, y un valor un poco más alto, en torno al 4-15% a nivel de proteína, ya que la tasa de error a nivel de proteína de cada uno de los experimentos está en torno a esos valores (Figura 81). Sin embargo, al integrarse todos los datos se observa un incremento significativo de la tasa de error global, sobre todo cuando los datos integrados son de una gran heterogeneidad, ya que experimentos de peor calidad penalizarán a los experimentos de mayor calidad. Dichos valores por tanto deben tener en cuenta relativamente, ya que no pretendemos obtener un valor global de identificaciones al integrar todos los datos, sino comparar cada uno de los experimentos por separado.

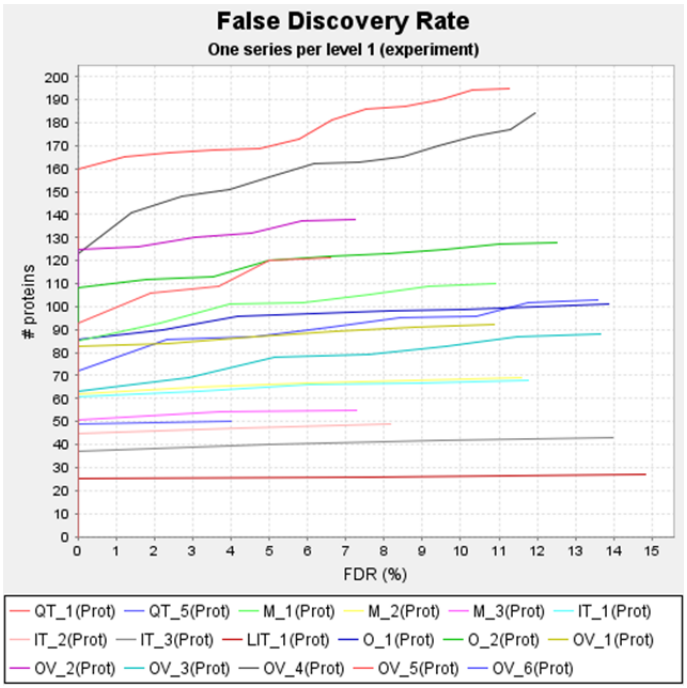


Figura 81. Curvas FDR de los datos re-analizados del PME6: Cada curva FDR representa el número de proteínas en función de la tasa de error, para un corte de un 1% a nivel de péptido. Se puede apreciar en el extremo derecho de las curvas que la tasa de error para ese corte oscila entre un 4% y un 15%.

Número de identificaciones y variabilidad entre réplicas

En la Figura 82 podemos ver el número medio de proteínas (A) y el número medio de péptidos y la desviación típica sobre las réplicas sin distinguir entre péptidos modificados y no modificados (B) o distinguiéndolos (C y D). El experimento QT_5 no muestra variabilidad ya que, como ya hemos comentado, únicamente realizó una réplica. Se puede ver cómo la variabilidad en general es baja tanto para péptidos como para proteínas, destacando la reproducibilidad de los datos de OV_5, el que además de tener un gran número de identificaciones, la variabilidad es mínima. En el otro extremo, los datos de M_1, OV_2, OV_3, OV_4 y OV_6, con una variabilidad mayor en el número de péptidos y proteínas.

También destacamos la disminución del número de péptidos y proteínas con respecto a los números enviados por parte del experimento LIT_1, que pasa de 293 a 22 péptidos y de 38 a 14 proteínas. En principio las búsquedas por parte del laboratorio LIT_1 se realizaron igualmente con Mascot, pero sí sabemos, que el experimento no salió bien. En este caso, con los datos re-analizados la diferencia con el resto de resultados es más acentuada.

Si comparamos estos números con los mostrados en los análisis propios de los participantes de la Figura 59, éstos cambian incrementándose unos o disminuyendo otros. Quedan patentes dos cosas: la variabilidad y la mayor cobertura de resultados que aporta el uso de diferentes herramientas de búsqueda, y la necesidad de un flujo común y centralizado de análisis para poder comparar de forma robusta los resultados obtenidos por el espectrómetro, controlando por tanto la variabilidad de los análisis de los datos.

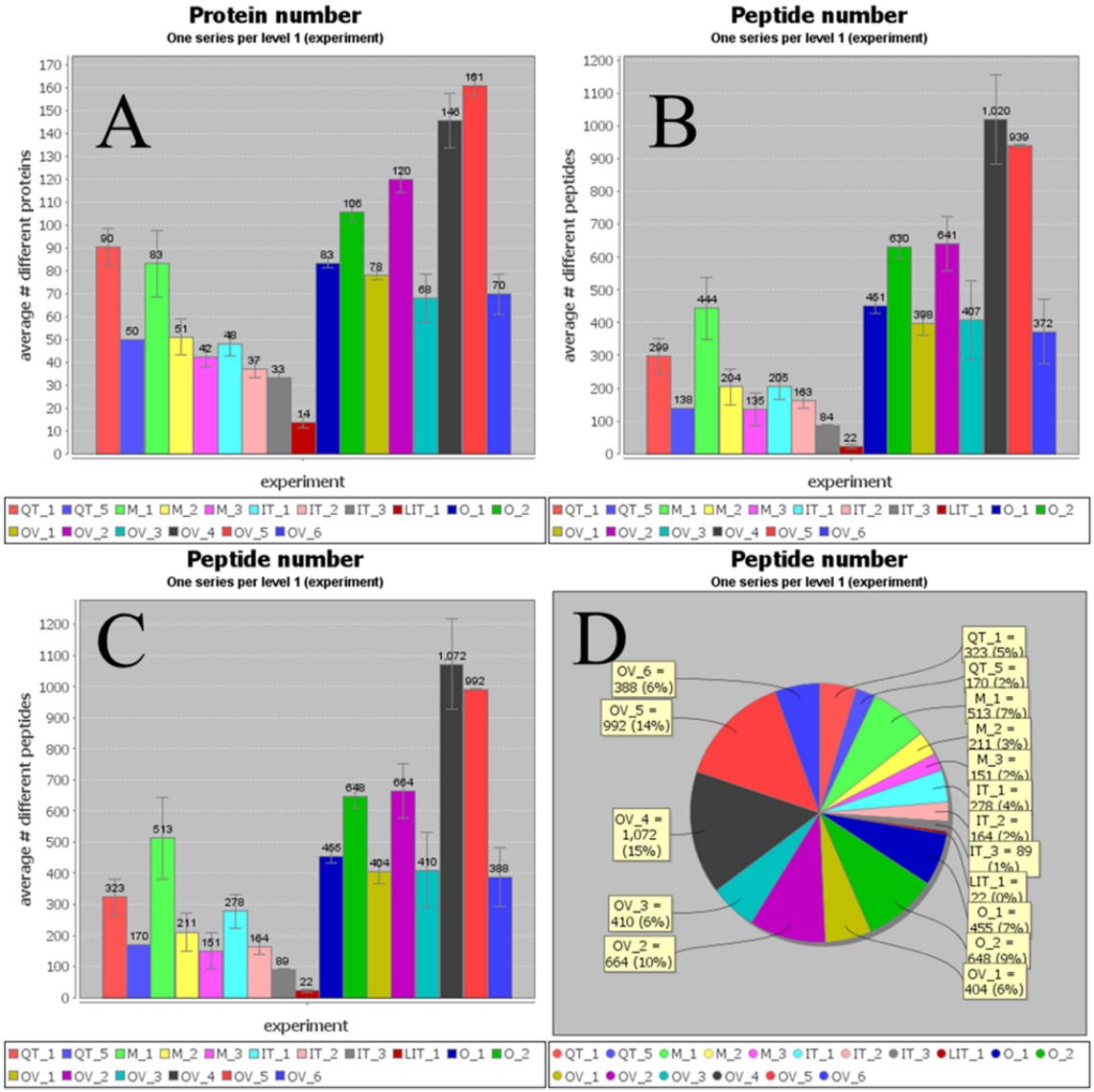


Figura 82. Número medio y desviación estándar de identificaciones de proteínas y péptidos: Para cada experimento se muestra el número medio de péptidos y proteínas obtenidos sobre las tres réplicas o sobre los diferentes análisis realizados sobre las 3 réplicas. En A se muestran el número medio de proteínas diferentes para cada experimento. En B se muestran el número medio de péptidos diferentes para cada experimento, tomando los péptidos modificados y no modificados como péptidos iguales. En C se muestran el número medio de péptidos diferentes para cada experimento distinguiendo entre péptidos modificados o no modificados. En D se muestra lo mismo que en C, pero en un gráfico de tarta.

Comparación de puntuaciones

Al igual que hemos hecho anteriormente, en la Figura 83 mostramos la comparación de las puntuaciones de los péptidos. En A, vemos superpuestas las distribuciones de puntuaciones Mascot:score de todos los experimentos. Destacan OV_2 (rosa oscuro) y OV_5 (rojo) como las distribuciones más altas y desplazadas a la derecha (puntuaciones más altas), y O_2 (verde oscuro) con un mayor número de mejores puntuaciones (mayores que 130) como se puede ver

en el recuadro interior. En B y C, representamos los valores de puntuación de un mismo péptido comparado en diferentes réplicas de un mismo experimento, dos a dos. En B, el experimento LIT_1, vemos un claro ejemplo de un problema en el experimento, ya que no se ven a penas puntos, es decir, no hay apenas solapamiento entre las réplicas. En C, vemos el caso contrario del experimento O_2, en el que las comparaciones dos a dos de las tres réplicas se ajustan bastante bien a la diagonal.

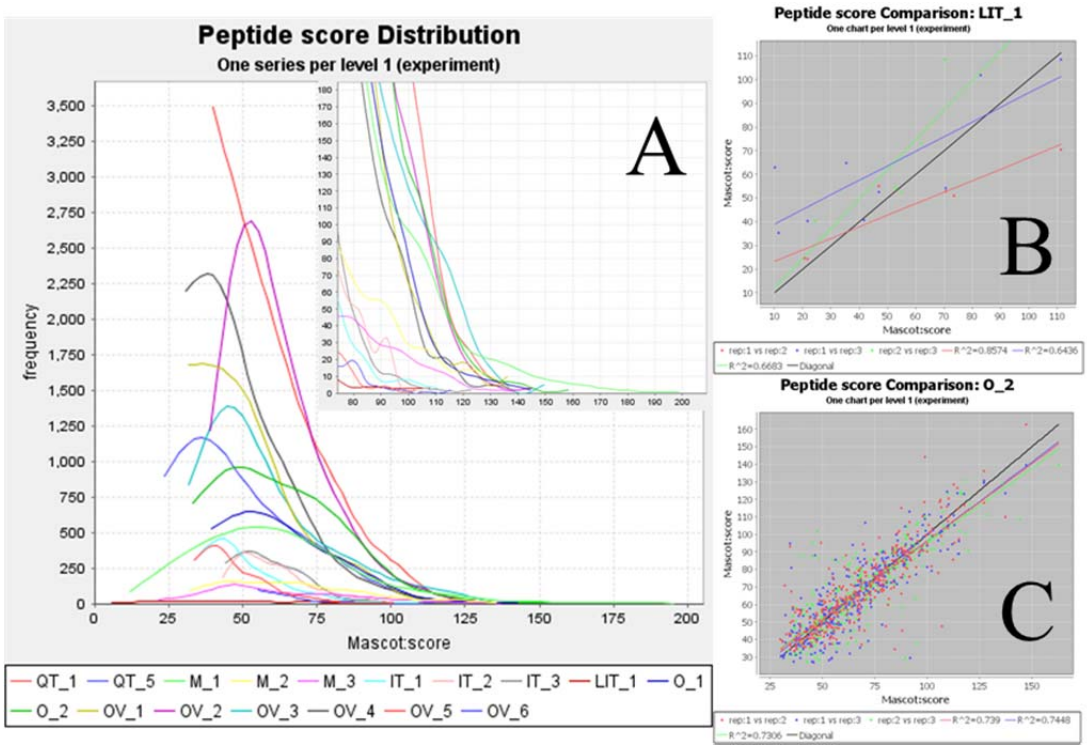


Figura 83. Comparación de puntuaciones: En A vemos las distribuciones de puntuación "*Mascot:score*" de todos los experimentos. En B la comparación de puntuaciones de péptidos en común entre las réplicas del experimento LIT_1. En C, lo mismo que en B, pero del experimento O_2.

Solapamientos

En la Figura 84 vemos los solapamientos a nivel de péptido entre las 3 réplicas de cada uno de los experimentos (exceptuando QT_2). Podemos observar que exceptuando los problemas ya comentados en el experimento LIT_1, aparentemente todos parecen tener un comportamiento parecido en cuestión de solapamiento.

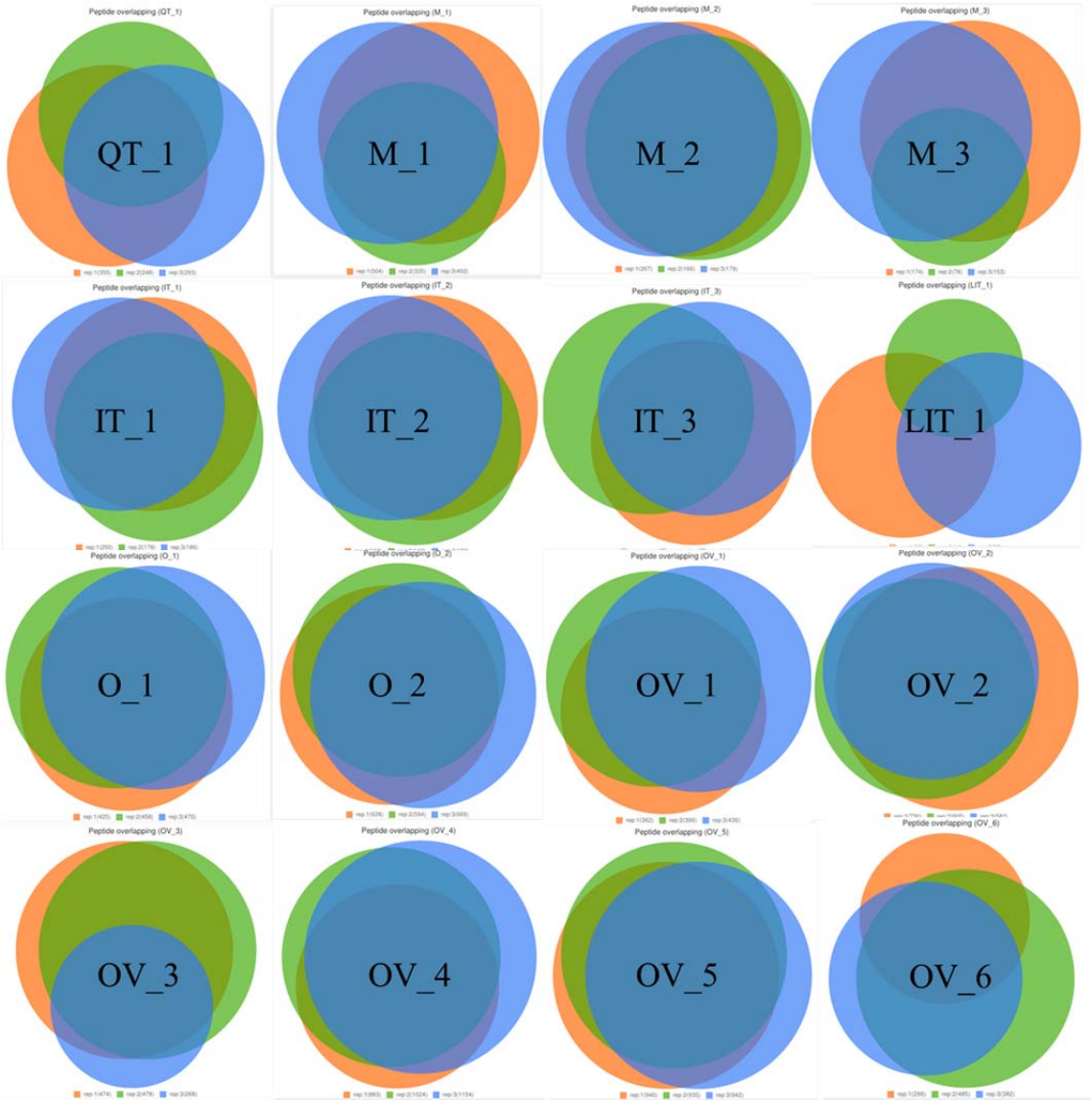


Figura 84. Solapamiento de péptidos entre las tres réplicas de cada experimento PME6: Réplica 1 (naranja), réplica 2 (verde), réplica 3 (azul).

Sin embargo, la herramienta MIAPE Extractor, además de mostrar los diagramas de Venn, muestra un resumen numérico de los solapamientos, como se mostró en la Figura 65. En la Tabla 8 hemos resumido los datos numéricos del solapamiento de todos los experimentos y podemos ver cómo los Orbitrap (salvo OV_1 y OV_6) parecen tener un grado superior de solapamiento (~50-60%) entre las réplicas, manteniéndose a pesar de tener un mayor número de identificaciones (~600-1300 péptidos en la unión). Cabe destacar por otro lado, el rendimiento del M_1 (MALDI TOF-TOF), con un solapamiento de un 40.4% en un total de 644 péptidos.

	QT_1	M_1	M_2	M_3	IT_1	IT_2	IT_3	LIT_1	O_1	O_2	OV_1	OV_2	OV_3	OV_4	OV_5	OV_6
Unión	490	644	294	219	306	232	125	45	600	807	553	815	570	1328	1190	575
Solapamiento total	131	260	133	56	118	107	49	4	324	462	258	500	211	729	725	181
Solapamiento total (%)	26,7%	40,4%	45.2%	25.6%	38.6%	46.1%	39.2%	8.9%	54.0%	57.2%	46.7%	61.3%	37.0%	54.9%	60.9%	31.5%

Tabla 8. Unión y solapamiento total entre réplicas: Se muestra el número de péptidos diferentes resultado de la unión de las tres réplicas para cada experimento, el número de péptidos diferentes que están en las tres réplicas, y el porcentaje de dicho conjunto de péptidos sobre la unión.

Reproducibilidad inter-laboratorio

Las gráficas de heat map de péptidos y proteínas de la Figura 85 muestran la reproducibilidad entre los distintos experimentos. A nivel de proteína (C) se ve una mayor reproducibilidad entre los experimentos con Orbitrap (excepto OV_3), junto con QT_1 y seguidos por M_1.

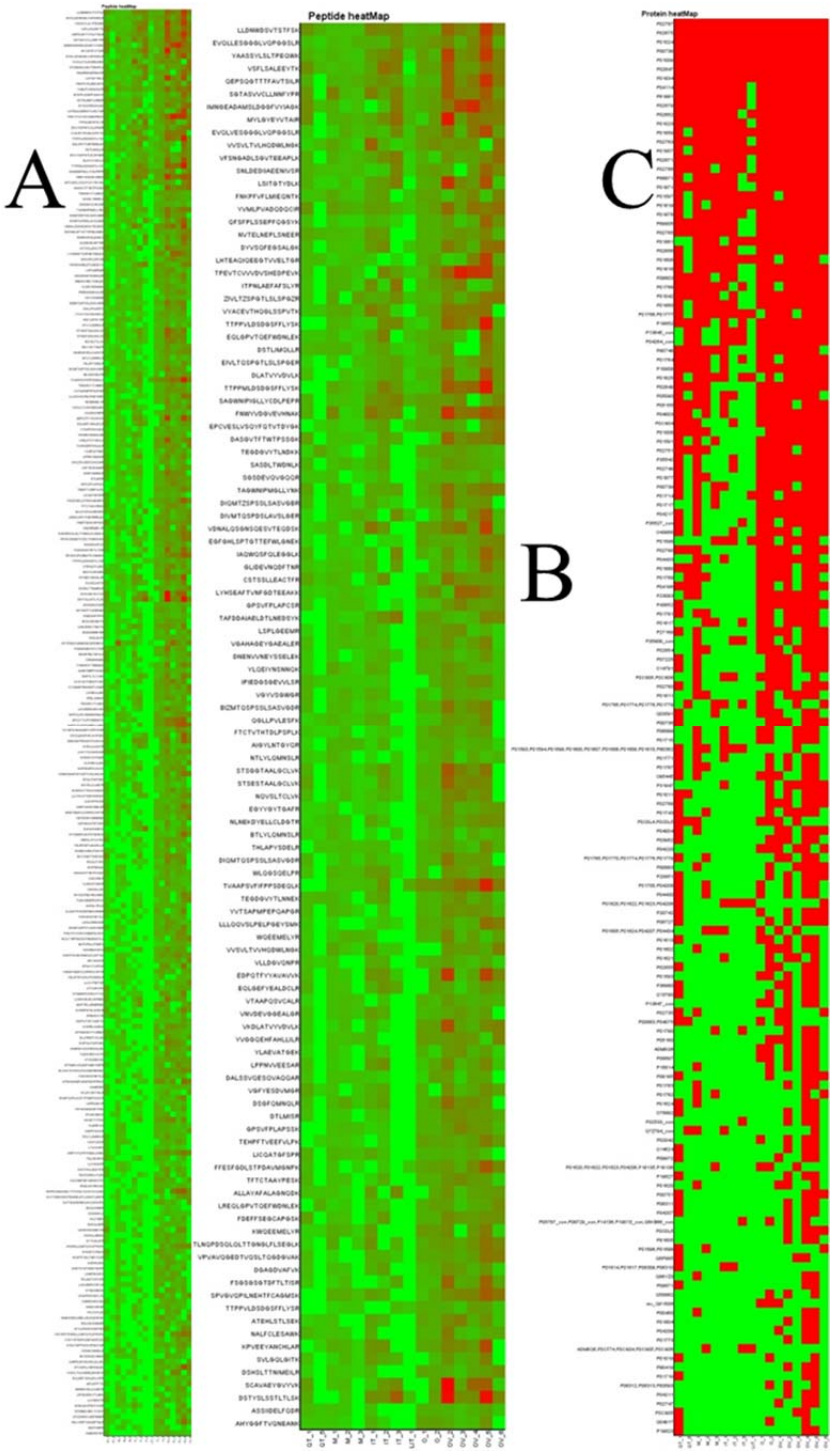


Figura 85. Heat maps de péptidos y proteínas: La escala de colores representa el número de veces que se ha visto un péptido o una proteína (en orden ascendente de verde a rojo). En el caso del heat map de péptidos se muestran los péptidos que se han visto hasta en 9 experimentos en A y hasta en 12 experimentos en B. La escala de colores es logarítmica en los dos casos. El experimento OV_1 fue omitido para no enmascarar los valores del resto de experimentos debido a un péptido sobre-representado en dicho experimento. En C, se muestran las proteínas que se han visto hasta en 3 experimentos. El orden de las columnas, ya que la leyenda de cada una es muy pequeña es, de izquierda a derecha: QT_1, QT_6, M_1, M_2, M_3, IT_1, IT_2, LIT_1, O_1, O_2, OV_2, OV_3, OV_4, OV_5 y OV_6.

Si miramos las proteínas exclusivas de cada experimento (Figura 86), es decir, las proteínas que únicamente se han detectado en un experimento, podemos observar en este caso que el experimento OV_5 tiene el mayor número de proteínas exclusivas. Además, si comparamos estos datos con los en la Figura 68 que mostraba los mismos datos para los resultados enviados por cada laboratorio, vemos que el número de proteínas exclusivas en general se ha reducido bastante. Por ejemplo, OV_4 tenía 137 y ahora sólo 29; OV_3 tenía 63 y ahora 15, etc… Esta disminución de las proteínas exclusivas se puede explicar por dos razones: 1) al re-analizar los datos con un flujo de trabajo común la variabilidad baja, ya que las condiciones del análisis son comunes a todos los experimentos; 2) por otro lado, al utilizar una única herramienta bioinformática (Mascot) la variabilidad en las identificaciones disminuye en comparación a la integración de los resultados de diferentes motores de búsqueda o herramientas de análisis. En conclusión, vemos que salvo los experimentos OV_4 y OV_5, las proteínas exclusivas aportadas por cada aproximación no es mayoritaria en los Orbitrap como vimos con los datos enviados por cada participante, ya que el número de proteínas exclusivas de los experimentos QT_1, M_1 o IT_1 es similar al de los Orbitrap O_1, O_2, OV_2, OV_3 y OV_6 e incluso superior que el de OV_1. Esto nos lleva a pensar que para sacar el máximo partido a los datos de los Orbitrap, Mascot no es la mejor herramienta de búsqueda. Como podemos ver en la Tabla 7, todos los participantes con Orbitrap excepto uno (O_2), utilizaron herramientas diferentes a Mascot.

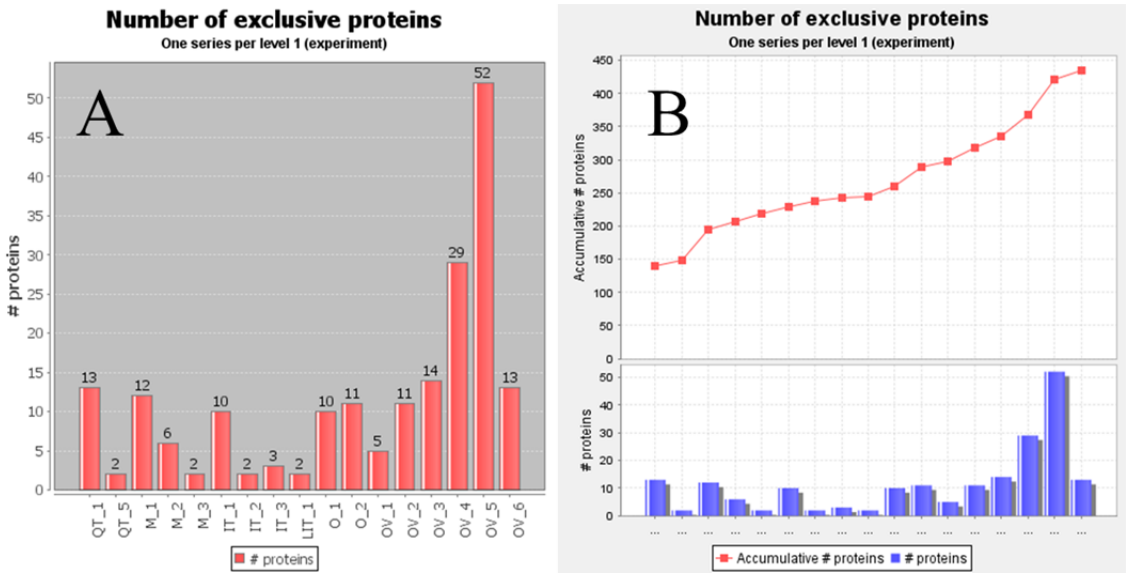


Figura 86. Proteínas exclusivas: En A se muestra el número de proteínas exclusivas de cada uno de los experimentos, es decir, que sólo se han visto en cada uno de los experimentos. En B, vemos además el número acumulativo de proteínas haciendo la unión secuencial de las proteínas de cada experimento en el orden mostrado.

Rendimiento en la digestión

Seleccionando los mismos experimentos que mostramos en la comparativa de los resultados enviados por los participantes (Figura 70 B), ahora vemos en la Figura 87 cómo el cambio más significativo se encuentra en el caso del experimento LIT_1, en el que según los datos enviados por el participante el 53% de los péptidos tenían una longitud mayor de 20, y sin embargo en los datos del re-análisis vemos que ese porcentaje baja hasta el 20%. También vimos que la digestión en LIT_1 no tuvo un gran rendimiento ya que había puntos de corte ausentes (*missed-cleavages*) en cerca del 50% de los péptidos. Es por ello por lo que la longitud de los mismos era mayor. Sin embargo, en los parámetros de búsqueda utilizados en el re-análisis de LIT_1, no se tuvieron en cuenta más que un máximo de 1 punto de corte de tripsina ausente, por lo que todos los péptidos con 2 o más lisinas o argininas en medio de la cadena peptídica no se detectaron y por eso también el número de péptidos identificados bajó de 61 a 22 y el número de proteínas, de 23 a 14. De esta manera hemos detectados una posible optimización de los resultados de LIT_1 utilizando un parámetros de número de puntos de corte ausentes de 2 o 3.

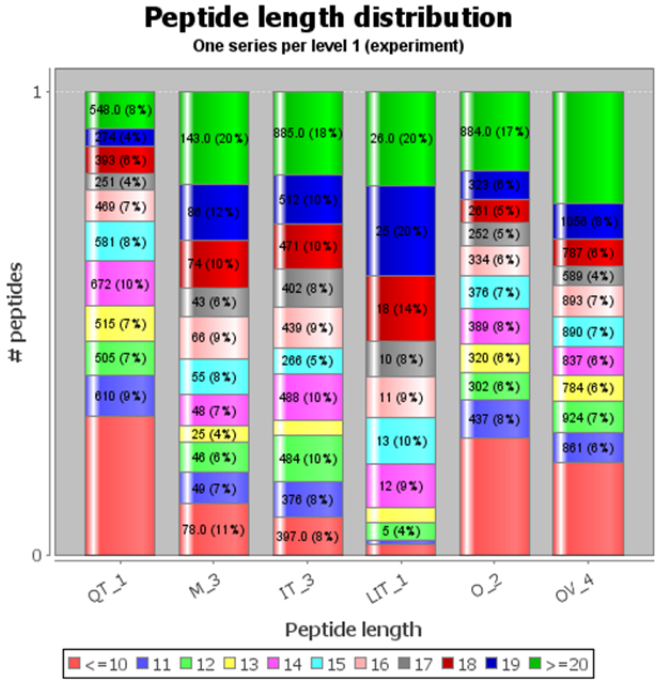


Figura 87. Distribución de longitudes de péptidos: Se muestra el número y porcentaje de péptidos con longitud menor o igual de 10 aminoácidos, 11, 12,…, 19 y 20 o más aminoácidos para los 6 experimentos seleccionados.

Cobertura de secuencia de proteínas

En cuanto a la cobertura de secuencia podemos ver en la Figura 88 cómo de nuevo los Orbitrap obtienen una cobertura media mayor, entre 18% y 25%, junto con los experimentos M_1, M_2 y IT_2. Los experimentos con menor cobertura de secuencia son QT_5 (sólo obtuvo una réplica), IT_3 y LIT_1.

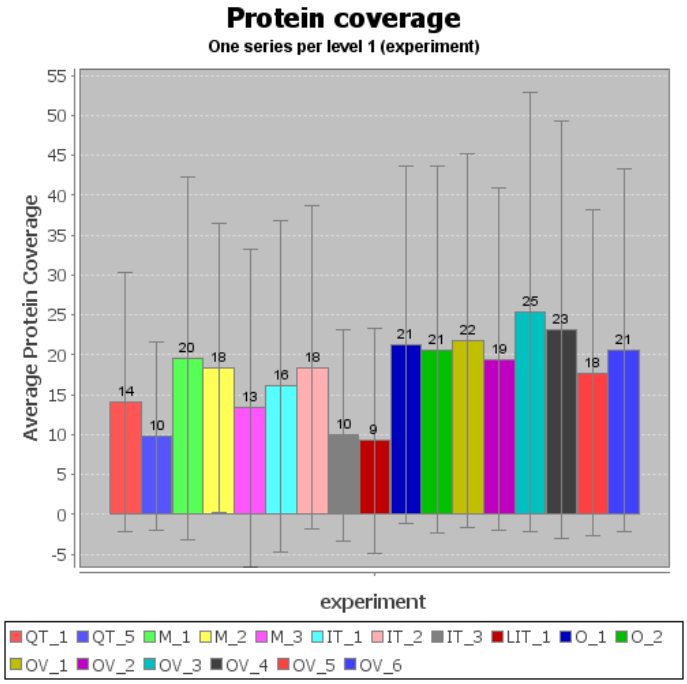


Figura 88. Cobertura de secuencia media y desviación típica de los datos re-analizados del PME6.

Redundancia en identificaciones

En cuanto a las veces que se ha visto cada secuencia peptídica, vemos en la Figura 89 (A) la proporción de péptidos identificados detectados una vez, dos veces,…, hasta 7 veces o más. Se observa cómo los Orbitrap, seguidos de las trampas iónicas (IT_x) detectan más veces cada secuencia peptídica, lo que ayudará a su identificación final. En la Figura 89 (B) vemos el porcentaje de péptidos vistos en una réplica (rojo), en dos (azul) o en las tres (verde). En este caso, igualmente los Orbitrap parecen tener mayor reproducibilidad sobre sus réplicas, seguidos en este caso de las trampas iónicas y los MALDI. El experimento LIT_1 obtuvo el menor resultado, ya que el 64% de sus péptidos fueron identificados únicamente en una réplica y no en las otras dos.

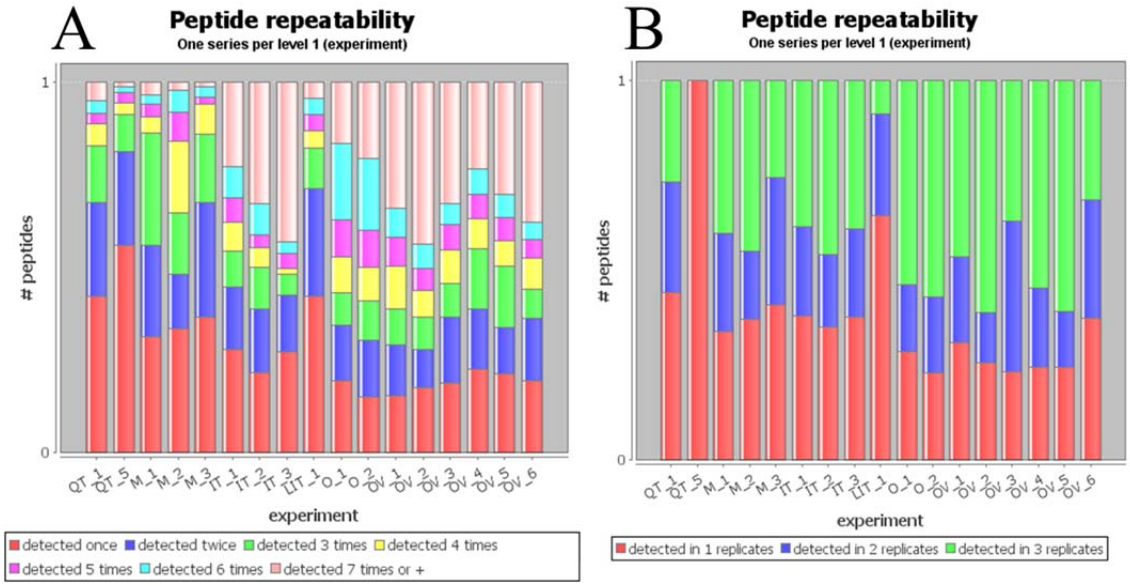


Figura 89. Repetitividad de detección de péptidos: En A se muestra el porcentaje (las barras están normalizadas) de péptidos identificados una vez (rojo), dos veces (azul), tres (verde), cuatro (amarillo), cinco (rosa oscuro), seis (azul claro) o siete o más veces (rosa claro). En B se muestra el porcentaje de péptidos identificados en una sola réplica (rojo), en dos réplicas (azul) o en tres réplicas (verde).

Sensibilidad para detectar las proteínas "spiked"

Al igual que hicimos en la anterior comparativa, incluyendo un filtro por los códigos de acceso de las 4 proteínas añadidas a la muestra (P61981 (30 µg), P00883 (3 µg), P02666 (0,3 µg) y P00489 (0,03 µg)), obtenemos la Figura 90 donde vemos que la proteína más abundante, la P61981 (Proteína 14-3-3 gamma) es detectada por todos los participantes. La siguiente proteína en abundancia, la P00883 (Aldolasa fructosa bifosfato A) es detectada por 11 de los 17 participantes (en los datos enviados por los participantes sólo 9 de 17 detectaron tal proteína). Luego, la proteína P02666 (Beta caseína) únicamente es detectada por el laboratorio M_1. Si relajamos el filtro por FDR (Figura 90 B), entonces 3 laboratorios la detectan (QT_5, M_1 y OV_2), en contraposición de los datos enviados por los participantes en los cuales los laboratorios M_3, OV_3 y OV_4 la detectaron con un 1% FDR de péptido. En el caso de la proteína presente en menor concentración, la P00489 (Glicógeno fosforilasa) es detectada por el laboratorio OV_4 y aun relajando el corte por FDR ningún otro la pudo detectar.

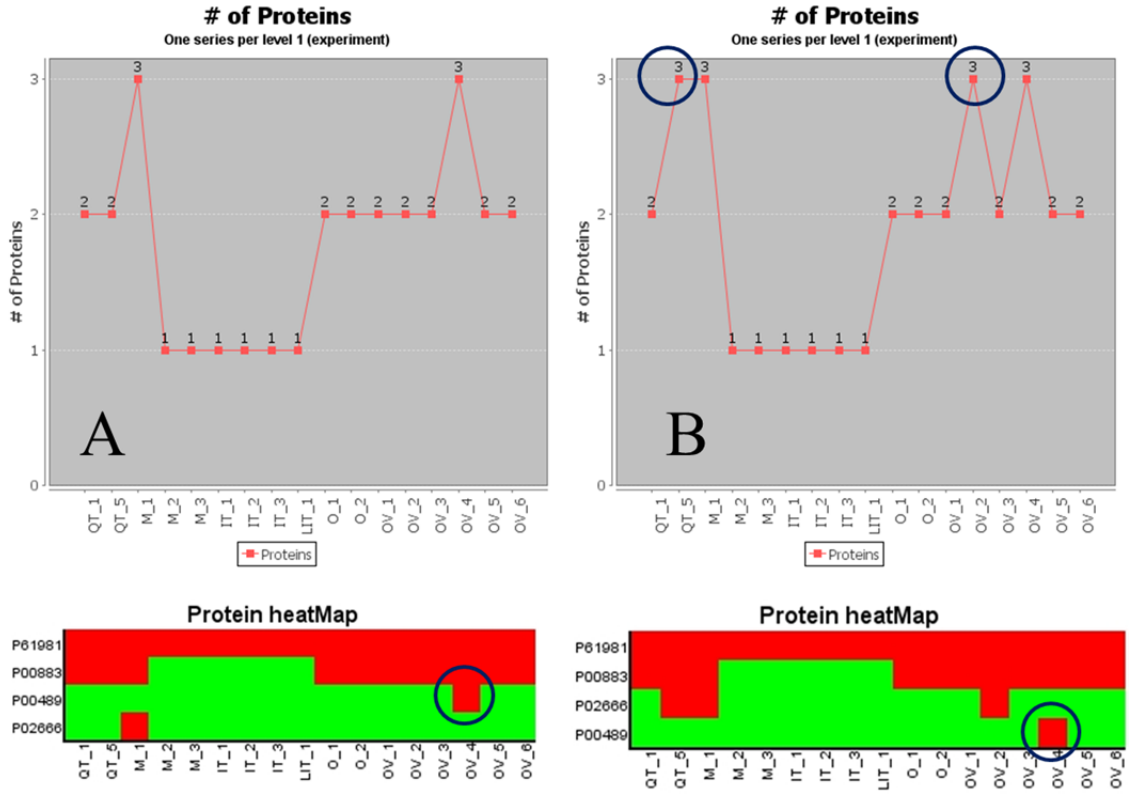


Figura 90. Análisis de las 4 proteínas añadidas a la muestra a distintas concentraciones: Se muestran el número de proteínas detectadas por cada participante (arriba), y el heat-map para las 3 proteínas detectadas (abajo). En A se muestran los datos tras aplicar los filtros descritos anteriormente, esto es, 1% FDR a nivel de péptido y en B con un filtro más laxo de un 5% de FDR a nivel de péptido.

Agrupamiento de proteínas

En el caso del re-análisis se ha podido realizar el agrupamiento de las proteínas cosa que no se pudo hacer en el caso del análisis de los datos enviados por los participantes, ya que en la mayoría de los casos enviaron proteínas sueltas, sin agrupar dado el caso, con otras.

Como hemos visto en este trabajo, la herramienta MIAPE Extractor implementa el algoritmo PAnalyzer (Prieto, Aloria et al. 2012) para reagrupar las proteínas según los péptidos que tengan compartidos con otras y clasificando dichas proteínas o grupos en 4 tipos: conclusivas, ambiguas, indistinguibles y no conclusivas. En la Figura 90 vemos la proporción de los distintos tipos de grupos o proteínas que el algoritmo de agrupamiento es capaz de distinguir. Las proteínas no conclusivas (amarillo) son proteínas que no se cuentan en el número final mostrado por la herramienta, siguiendo las directrices del trabajo de Aebersold y Nesvizhskii (Nesvizhskii y Aebersold 2005), ya que son proteínas cuyos péptidos detectados están explicados por proteínas que realmente están en la muestra y por tanto pueden estar presentes o no.

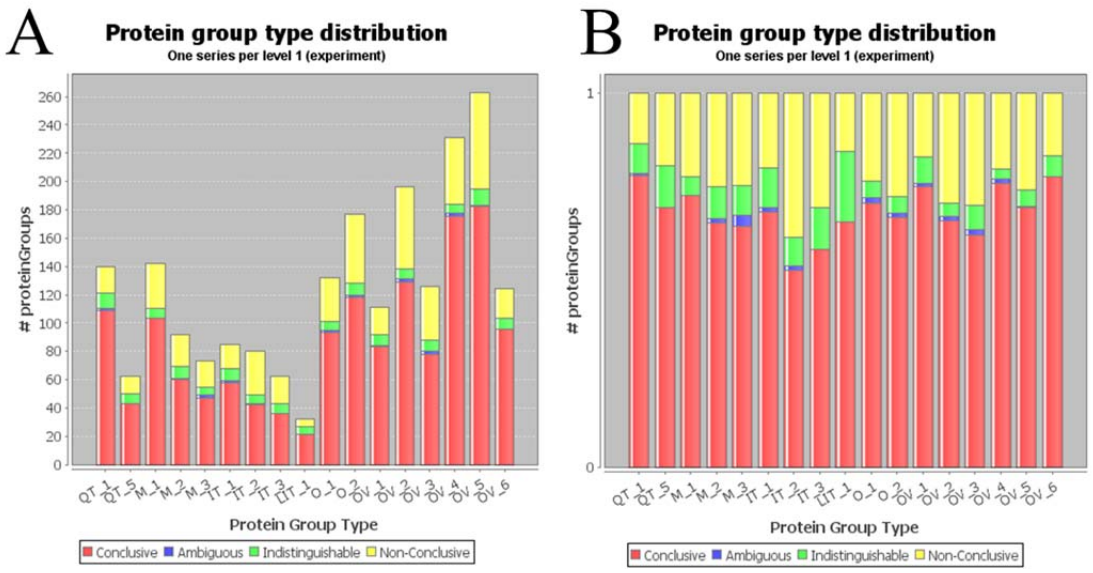


Figura 91. Distribución de la clasificación del agrupamiento del algoritmo PAnalyzer: En rojo se muestran las proteínas conclusivas, en azul los grupos ambiguos, en verde los grupos indistinguibles y en amarillo las proteínas no conclusivas. En A se muestran los números absolutos y en B los números normalizados.

Si nos fijamos en 3 de las proteínas añadidas a la muestra (*spiked*), vemos que en varios casos se identificaron dentro de grupos indistinguibles, es decir que, no se pudo encontrar el/los péptido/s necesarios como para identificar de manera indistinguible la proteína añadida a la muestra:

- El laboratorio LIT_1 detecta la proteína más abundante de las 4, la P61981 (14-3-3 protein gamma) pero dentro de un grupo indistinguible compuesto por 7 proteínas 14-3-3 (Figura 92).

14-3-3 protein theta OS=Homo sapiens GN=YWHAQ PE=1 SV=1
14-3-3 protein beta/alpha OS=Homo sapiens GN=YWHAB PE=1 SV=3
14-3-3 protein sigma OS=Homo sapiens GN=SFN PE=1 SV=1
14-3-3 protein gamma OS=Homo sapiens GN=YWHAG PE=1 SV=2
14-3-3 protein epsilon OS=Homo sapiens GN=YWHAE PE=1 SV=1
14-3-3 protein zeta/delta OS=Homo sapiens GN=YWHAZ PE=1 SV=1
14-3-3 protein eta OS=Homo sapiens GN=YWHAH PE=1 SV=4

Figura 92. Grupo indistinguible detectado por LIT_1 con la proteína P61981 (14-3-3 protein gamma).

- El laboratorio QT_5 detectan la segunda proteína en abundancia, la proteína P00883 (Aldolasa fructosa bifosfato A, de conejo) dentro también de un grupo indistinguible en el que también está la misma proteína de la especie humana P04075 (Figura 93).

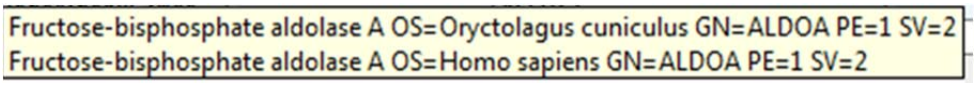


Figura 93. Grupo indistinguible detectado por QT_5 con la proteína P00883 (Aldolasa fructosa bifosfato A, de conejo).

- Por su parte, el laboratorio OV_4, el único que detectó la proteína menos abundante, es decir, la proteína P00489 (Glicógeno fosforilasa de conejo), la detectó también dentro de un grupo indistinguible de proteínas que compartían un único péptido. En este caso las otras dos proteínas eran la misma proteína, una de músculo humano y otra de cerebro humano.

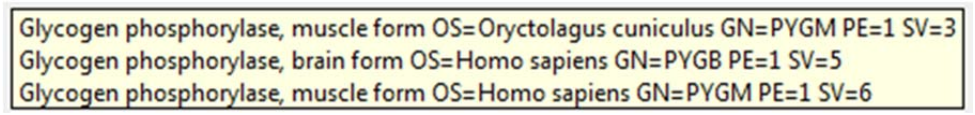


Figura 94. Grupo indistinguible detectado por OV_4 con la proteína P00489 (Glicógeno fosforilasa de conejo).

Todos estos análisis se pueden hacer de forma muy fácil y rápida con la herramienta MIAPE Extractor pudiendo hacer filtros en la propia tabla de identificaciones y pudiendo ver los agrupamientos de las proteínas resultantes del algoritmo PAnalyzer.

Sensibilidad y precisión

En este caso, al igual que en la comparativa de los resultados obtenidos por cada laboratorio, consideramos las proteínas que se han visto en al menos dos laboratorios como las proteínas verdaderas positivas (TP) para hacer los cálculos de sensibilidad, especificidad y precisión. Así pues, exportando las proteínas y cogiendo las detectadas en dos o más laboratorios tenemos un total de 196 proteínas que consideraremos verdaderas positivas.

La herramienta nos muestra la Figura 95 en la que vemos que la sensibilidad, es decir, la fracción de verdaderos positivos identificados de los 196 existentes, está por encima del 90% en la mayoría de los caso, excepto en el caso de QT_1, QT5, IT_2 e IT_3. En cuanto a la precisión, es decir, el número de identificaciones verdaderamente correctas dentro de las identificaciones que pasan el corte, los valores son bastante homogéneos, destacando sorprendentemente los valores ligeramente peores de los Orbitrap Velos, por debajo del 90%.

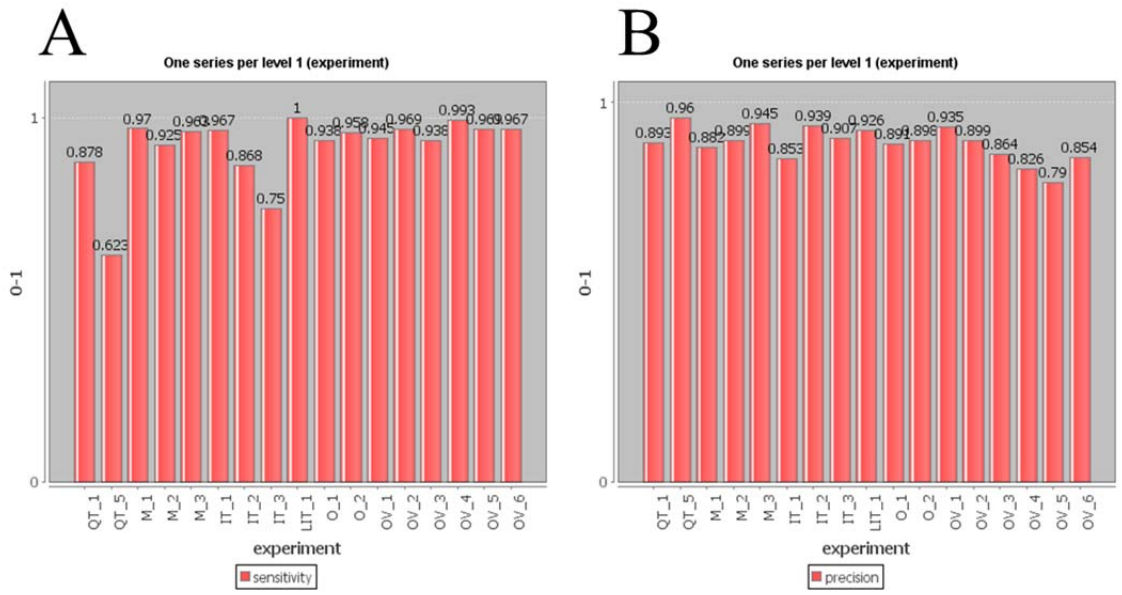


Figura 95. Sensibilidad y precisión del re-análisis de los datos de los participantes del PME6. Considerando como verdaderas positivas las 196 proteínas que se vieron en al menos dos laboratorios, mostramos la sensibilidad (A) y la precisión (B) de cada uno de los participantes.

Nube de proteínas

Una de las representaciones más vistosas de la herramienta es la llamada nube de proteínas, que recoge las palabras de las descripciones de todas las proteínas en el proyecto de inspección y las representa en forma de nube, con diferentes tamaños según el número de veces en el que ocurren. La herramienta permite ignorar palabras concretas (ej: "*gene*", "*decoy*", "*human*",…) o mostrar únicamente las palabras con una longitud mínima. Así pues, estas representaciones nos dan una idea sobre el tipo de proteínas contenidas en la muestra. La Figura 96 nos muestra tres nubes diferentes generadas con las proteínas del experimento PME6.

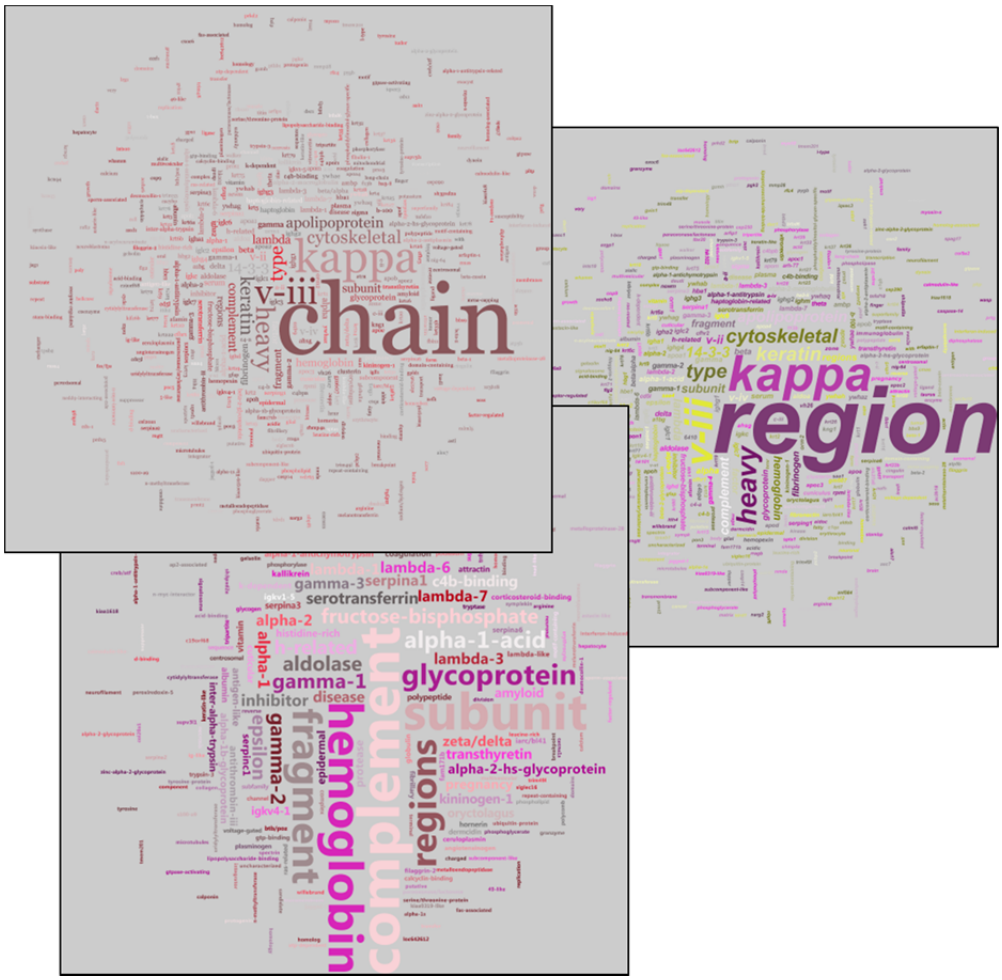


Figura 96. Ejemplos de nubes de proteínas. En este caso se generaron tres nubes de proteínas, aplicando diferentes filtros de longitud y de palabras.

Envío a ProteomeXchange

Como hemos comentado en la sección de resultados (4.2.5), una de las principales características de la herramienta MIAPE Extractor es que, además de permitir la agregación, análisis y comparación de grandes cantidades de datos, permite la preparación de los ficheros necesarios para el envío de los datos a ProteomeXchange. Así pues, simplemente pinchando en el botón "ProteomeXchange" de la interfaz de inspección de datos, se muestra una pantalla para introducir los metadatos necesarios para el envío (título y descripción del experimento o de los datos y una lista de palabras clave relacionadas).

Como último paso necesario para completar el envío por parte del usuario, sería seleccionar los ficheros de datos crudos y añadirlos en el árbol resumen del envío, asociándolos con los nodos correspondientes. En el caso de los datos re-analizados del experimento PME6, debido a que se utilizó como plantilla de metadatos MS los mismos documentos MIAPE MS que en su

día se hicieron a mano en la herramienta online generadora de documentos MIAPE, y en éstos ya se puso un link al fichero de datos crudos correspondiente alojado en nuestro servidor FTP (en la sección "*Resulting Data*"), la herramienta detectará y validará ese link y añadirá automáticamente dichos ficheros al árbol resumen del envío.

Finalmente, se realizó el envío de los datos a ProteomeXchange, incluyendo un total de 262 ficheros: 17 ficheros PRIDE XML, 49 ficheros de datos crudos, 49 ficheros mzIdentML, 43 ficheros mzML, 6 ficheros mgf y 98 informes MIAPE, siendo más de 28 Gbytes de información, estando la mayor parte de los ficheros comprimidos. Sin embargo, en el momento de la escritura de esta tesis el conjunto de datos aún no había sido validado por el equipo PRIDE-EBI/EMBL y por tanto no podemos escribir aquí el identificador PX asignado.

MIAPE section	Item	Concept	XPath (under /mzML/)	CV term accession	CV preferred name	Notes	Allow children
1. General Features							
1.1 Global descriptors	Responsible person	contact name	fileDescription/contact/cvParam/@accession	MS:1000586	contact name		no
		contact organization	fileDescription/contact/cvParam/@accession	MS:1000590	contact organization		no
		contact email	fileDescription/contact/cvParam/@accession	MS:1000589	contact email		no
		contact role	fileDescription/contact/cvParam/@accession	MS:1001266	role type		yes
	Instrument manufacturer, model	instrument manufacturer	instrumentConfigurationList/instrumentConfiguration /cvParam/@accession	MS:1000031	instrument model	Direct children of (MS:1000031) describe the manufacturers, and grandchildren describe the models of the instruments.	yes
		instrument model	instrumentConfigurationList/instrumentConfiguration /cvParam/@accession				yes
	Customizations (summary)	mass spectrometer customizations	instrumentConfigurationList/instrumentConfiguration /cvParam/@accession	MS:1000032	customization	Check value is not empty	no
2. Ion sources							
2.1 Electrospray ionization (if 2.0 is	Ion source	Ion source	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1000073	electrospray ionization		yes
	Supply type (static, or fed)	Supply type	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1001941	electrospray supply type		yes

MS:000073)	Interface manufacturer, model	Interface manufacturer	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1001953	source interface manufacturer	Check value is not empty	no
		Interface model	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1001932	source interface model	Check value is not empty	no
	Sprayer type, manufacturer, model	Sprayer type	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1001934	source sprayer type		yes
		Sprayer manufacturer	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1001935	source sprayer manufacturer	Check value is not empty	no
		Sprayer model	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1001936	source sprayer model	Check value is not empty	no
	Other parameters if discriminant for the experiment	Other parameters if discriminant for the experiment	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	-	-	cannot be validated	-
2.2 MALDI (if 2.0 is MS:000075)	Ion source	Ion source	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1000075	matrix-assisted laser desorp. ioniz.		no
	Plate composition (or type)	Plate composition (or type)	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1001938	sample plate type		yes
	Matrix composition	Matrix composition	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1000834	matrix solution	Check value is not empty	no

	PSD (or LID/ISD) summary, if performed	PSD summary, if performed	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1000135	post-source decay	The absence of the term means "no"	At least one of these three options should be present	no
		LID summary, if performed	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1000075	matrix-assisted laser desorp. ioniz.			no
		ISD summary, if performed	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1001880	in-source collision-induced dissociation			no
	Operation with or without delayed extraction	Operation with or without delayed extraction	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1000246	delayed extraction	The absence of this term means "no delayed extraction"		no
	Laser type and wavelength (nm)	Laser type	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1000842	laser type			yes
		Laser wavelength (nm),	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1000843	laser wavelength	Check value is not empty		no
	Other laser related parameters, if discriminating for the experiment	Other laser related parameters, if discriminating for the experiment	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1000841	laser attribute	The absence of this term means no parameters		yes
2.3 Other ion source	Description of the ion source	Description of the ion source	instrumentConfigurationList/instrumentConfiguration /componentList/source/cvParam/@accession	MS:1000008	ionization type	Different from MALDI or ESI cv terms		yes
	Relevant parameters	Other source relevant parameter				Cannot be validated		

3. Post-source components							
3.1 Analyzer	Ion optics, 'simple' quadrupole, hexapole, Paul trap, linear trap, magnetic sector, FT-ICR, Orbitrap	Analyzer type	instrumentConfigurationList/instrumentConfiguration /componentList/analyzer/cvParam/@accession	MS:1000443	mass analyzer type		yes
	TOF drift tube	Reflectron status (on, off, none)	instrumentConfigurationList/instrumentConfiguration /componentList/analyzer/cvParam/@accession	MS:1000021	reflectron state	Just if mass analyzer type is TOF (MS:1000084)	yes
3.2 Activation / dissociation	Instrument component where the activation / dissociation occurs	Instrument component where the activation / dissociation occurs	To be decided			Cannot be validated	
	Gas type (when used)	Gas type	run/spectrumList/spectrum/precursorList/precursor/activation /cvParam/@accession	MS:1000419	collision gas	The absence of this term means, no gas used. If present, check value is not empty	no
	Activation / dissociation type	Activation / dissociation type	run/spectrumList/spectrum/precursorList/precursor/activation /cvParam/@accession	MS:1000044	dissociation method		yes
4. Spectrum and peak list generation and annotation							
4.1 Data Acquisition	Software name and version	Software name	softwareList/software/cvParam/@accession	MS:1001455	acquisition software		yes
		Software version	softwareList/software/@version	-	-	Mandatory attribute in mzML	-

	Acquisition parameters	Acquisition parameter file	fileDescription/sourceFileList/sourceFile/@name and @location referenced from scanSettingsList/scanSettings/sourceFileRefList/sourceFileRef/@ref	MS:1000740	parameter file	At least one of these information must be present. If present, check value is not empty	no
		Acquisition parameters explicit description	softwareList/software/cvParam/@accession	New term	acquisition parameters		
4.2 Data analysis	Software name and version	Software name	softwareList/software/cvParam/@accession	MS:1001457	data processing software		yes
		Software version	softwareList/software/@version	-	-	Mandatory attribute in mzML	-
	Parameters used in the generation of peak lists or processed spectra	Parameters used in the generation of peak lists or processed spectra	dataProcessingList/dataProcessing/processingMethod /cvParam/@accession	MS:1000630	data processing parameter		no
4.3 Resulting data for each dataset	Location of source ("raw") and processed files	Location of source ("raw") and processed files	fileDescription/sourceFileList/sourceFile/@name and @location	-	-	Mandatory attributes	-
	The chromatogram(s) for SRM data and other relevant cases. To validate this, we need to validate all the elements needed for a correctly reported	Chromatogram binary data array	run/chromatogramList/chromatogram/binaryDataArrayList /binaryDataArray/cvParam/@accession	MS:1000513	binary data array	Already in minimal semantic validation: no extra rules to add.	yes
		Chromatogram binary data type	run/chromatogramList/chromatogram/binaryDataArrayList /binaryDataArray/cvParam/@accession	MS:1000518	binary data type		yes
		Chromatogram binary data compression type	run/chromatogramList/chromatogram/binaryDataArrayList /binaryDataArray/cvParam/@accession	MS:1000572	binary data compression type		yes

	chromatogram	Chromatogram intensity array	run/chromatogramList/chromatogram/binaryDataArrayList /binaryDataArray/cvParam/@accession	MS:1000515	intensity array	Check that one term or other is present (not both)	no
		Chromatogram time array	run/chromatogramList/chromatogram/binaryDataArrayList /binaryDataArray/cvParam/@accession	MS:1000595	time array		no
4.3 Resulting data for each spectrum or peaklist	m/z and intensity values	m/z and intensity values	run/spectrumList/spectrum/binaryDataArrayList/binaryDataArray /cvParam/@accession	MS:1000515	Intensity array	Check that one term or the other is present (not both)	no
				MS:1000514	m/z array		
	MS level	MS level	run/spectrumList/spectrum/cvParam/@accession	MS:1000511	ms level	Check value is not empty	no
	Ion mode	Ion mode	run/spectrumList/spectrum/cvParam/@accession	MS:1000465	scan polarity		yes
	For MS level 2 and higher, precursor m/z and charge if known, with the full mass spectrum / peaklist containing that precursor peak, where available.	Precursor m/z	run/spectrumList/spectrum/precursorList/precursor/selectedIonList /selectedIon/cvParam/@accession	MS:1000744	selected ion m/z	Check value is not empty	no
		Precursor charge if known	run/spectrumList/spectrum/precursorList/precursor/selectedIonList /selectedIon/cvParam/@accession	MS:1000041	charge state	Check value is not empty	no
		Full mass spectrum / peaklist containing that precursor peak.	run/spectrumList/spectrum/precursorList/precursor/spectrumRef	no cv: xpath used instead	-	Check if MS level >= 2 that: spectrumRef or sourceFileRef and externalSpectrumID are not empty	-
			run/spectrumList/spectrum/precursorList/precursor/sourceFileRef and run/spectrumList/spectrum/precursorList/precursor/externalSpectrumID				

Tabla 9. Correspondencia entre la información requerida por las directrices MIAPE MS y los elementos del formato estándar mzML. Columnas: (1) sección MIAPE; (2) información requerida en esa sección; (3) explicación del concepto requerido; (4) ruta XPath al elemento del mzML donde debe estar anotada dicha información; (5) términos de la ontología que deben utilizarse para anotar esa información; (6) nombre de el/los término(s) de la anterior columna; (7) ciertos comentarios acerca de cómo validar la información; (8) se muestra si se permite el uso de los términos "hijos" del término de la quinta columna. Las celdas en rojo corresponden a los términos que se añadieron a la ontología PSI-MS tras este estudio.

MIAPE section	Item	Concept	XPath (under /MzIdentML/)	CV term accession	CV preferred name	Notes		Allow children
1. General Features								
1.1 Global descriptors	Responsible person	contact name	AuditCollection/Person/@name @lastName @firstName referenced by Provider/ContactRole/@contact_ref			Any of these attributes must be present	A valid contact name or a valid contact role will be enough	
		contact role	Provider/ContactRole/Role/cvParam/@accession	MS:1001266	contact role			no
		contact affiliation	AuditCollection/Organization/@name referenced by Provider/ContactRole/@contact_ref			The name attribute must be present		
		contact email	AuditCollection/Person/cvParam /@accession AuditCollection/Organization /cvParam/@accession AuditCollection/Organization /cvParam/@accession referenced by AuditCollection/Organization /Parent/@organization_ref	Referenced by Provider/ContactRole /@contact_ref	MS:1000589	contact email	This term must be present at least in one of the three elements	no
	Software package(s) - name, version and manufacturer	Software name	AnalysisSoftwareList/AnalysisSoftware/SoftwareName /cvParam/@accession		MS:1001456	analysis software		
		Software version	AnalysisSoftwareList/AnalysisSoftware/@version			This attribute must be present		

		Software manufacturer	AnalysisSoftwareList/AnalysisSoftware/ContactRole/Role /cvParam/@accession	MS:1001267	software vendor	The value must not be empty	no
	Customizations made to that software	Software customizations	AnalysisSoftwareList/AnalysisSoftware/Customizations			The absence of this element means no customizations	
	Availability of that software	Availabilty of the software if publicly available	AnalysisSoftwareList/AnalysisSoftware/@uri			The absence of this attribute means no publicly available	
	Location of the files generated by the procedure	Location of the files generated if made available in a public repository	DataCollection/Inputs/SourceFile/@location			The absence of this attribute means no public repository available	
2. Input data and parameters							
2.1 Input data	Description and type of the input MS data	Input data description	DataCollection/Inputs/SpectraData/@name			This attribute must be present	
		Input data type	DataCollection/Inputs/SpectraData/FileFormat/cvParam/@accession	MS:1000560	mass spectrometer file format		yes
	Location of the input MS data	Location of the input MS data	DataCollection/Inputs/SpectraData/@location			Mandatory attribute.	
2.2 Input parameters	Database(s) queried; description and version (including number of sequences searched). If decoy, method used to	Sequence Database(s) or spectrum library	DataCollection/Inputs/SearchDatabase/DatabaseName /cvParam/@accession	MS:1001013	database name		yes
		Database version	DataCollection/Inputs/SearchDatabase/@version			This attribute must be present	

	generate decoy sequences and whether it was concatenated / separated from the target database.	Number of sequences searched	DataCollection/Inputs/SearchDatabase/@numDatabaseSequences			This attribute must be present	
		Decoy database description	DataCollection/Inputs/SearchDatabase/cvParam/@accession	MS:1001450	decoy DB details	Just in case of using decoy database(s)	yes
	Taxonomical restrictions applied	Filter type	AnalysisProtocolCollection/SpectrumIdentificationProtocol /DatabaseFilters/Filter/filterType/cvParam/@accession	MS:1001511	Sequence database filter types	The absence of these terms means no filter in the database	yes
		Inclusions and exclusions in the database entries	AnalysisProtocolCollection/SpectrumIdentificationProtocol /DatabaseFilters/Filter/include/cvParam/@accesion *and* AnalysisProtocolCollection/SpectrumIdentificationProtocol /DatabaseFilters/Filter/exclude/cvParam/@accession	MS:1001512	Sequence database filters		yes
	Selected scoring scheme in the software	Selected scoring scheme in the software	AnalysisProtocolCollection/SpectrumIdentificationProtocol /AdditionalSearchParams/cvParam/@accession	MS:1001961	peptide spectrum match scoring algorithm		yes
	Specified cleavage agent(s). Cleavage agent rules defined by the user.	Specified cleavage agent(s)	AnalysisProtocolCollection/SpectrumIdentificationProtocol /Enzymes/Enzyme/EnzymeName/cvParam/@accession	MS:1001045	cleavage agent name	Check either a valid cvParam under enzyme name or a siteRegexp is provided	
		Cleavage agent rules defined by the user	AnalysisProtocolCollection/SpectrumIdentificationProtocol /Enzymes/Enzyme/SiteRegexp				
	Allowed number of missed cleavages	Allowed number of missed cleavages	AnalysisProtocolCollection/SpectrumIdentificationProtocol /Enzymes/Enzyme/@missedCleavages			The absence of this attribute means no enzyme digestion (following mzid spec. doc.)	

	Additional parameters related to cleavage	Additional parameters related to cleavage	AnalysisProtocolCollection/SpectrumIdentificationProtocol/Enzymes			No extra validation rules to add.	
	Permissible amino acids modifications	Permissible amino acids modifications	AnalysisProtocolCollection/SpectrumIdentificationProtocol /ModificationParams/SearchModification/cvParam/@accession	UNIMOD:0 - MOD:00000 - MS:1001471	unimod root term - protein modification - peptide modification details	If 'ModificationParams' element is present, at least one cvParam must be present	yes
	Precursor-ion and fragment ion mass tolerances for tandem MS (when applicable)	Precursor-ion and fragment ion mass tolerances for tandem MS	AnalysisProtocolCollection/SpectrumIdentificationProtocol /FragmentTolerance	MS:1001412	search tolerance plus value	Check if elements are present and contain both terms. For PMF, omit the FragmentTolerance element.	no
			AnalysisProtocolCollection/SpectrumIdentificationProtocol /ParentTolerance	MS:1001413	search tolerance minus value		
	Mass tolerance for PMF (when applicable)	Mass tolerance for PMF	AnalysisProtocolCollection/SpectrumIdentificationProtocol /ParentTolerance	MS:100141 - MS:1001413	search tolerance plus value - search tolerance minus value		no
	Any other relevant parameters		AnalysisProtocolCollection/SpectrumIdentificationProtocol /AdditionalSearchParams/cvParam/@accession			No additional rules to add	
3. The output from the procedure							
3.1 For identified	Accession code in the queried database	Protein accession code	SequenceCollection/DBSequence/@accession			Mandatory attribute	

proteins	Protein score(s)	Protein score(s)	DataCollection/AnalysisData/ProteinDetectionList /ProteinAmbiguityGroup/ProteinDetectionHypothesis /cvParam/@accession	MS:1001153	search engine specific score	Check that at least one child term of MS:1001153 or MS:1001116 has been provided		yes
				MS:1001116	single protein result details			yes
	Number of different peptide sequences (without considering modifications or charge state) assigned to the protein					Implicit in mzIdentML schema		
	Identity of supporting peptides		DataCollection/AnalysisData/ProteinDetectionList /ProteinAmbiguityGroup/ProteinDetectionHypothesis /PeptideHypothesis			Implicit in mzIdentML schema		
	In the case of PMF, number of matched / unmatched peaks	Number of matched / unmatched peaks	DataCollection/AnalysisData/ProteinDetectionList /ProteinAmbiguityGroup/ProteinDetectionHypothesis /cvParam/@accession	MS:1001121	number of matched peaks	Just in case of PMF	This term must be present	no
				MS:1001362 or MS:1001124	number of unmatched peaks or number of peaks submitted		At least one of these two terms must be present	no
3.2 For identified peptides	Sequence	Sequence	SequenceCollection/Peptide/PeptideSequence			Mandatory element		
	Peptide score(s)	Peptide score(s)	DataCollection/AnalysisData/SpectrumIdentificationList /SpectrumIdentificationResult/SpectrumIdentificationItem /cvParam/@accession	MS:1001143	search engine specific score for peptides	At least one child of one of these two terms must be present.		yes

				MS:1001105	peptide result details		yes
	Chemical modifications (induced by experimental conditions) and modifications of biological source (naturally-occurring); amino acid sequence polymorphisms	Peptide modifications	SequenceCollection/Peptide/Modification			No additional rules to add.	
	Evidence for the presence and location of the modifications		SequenceCollection/Peptide/Modification/cvParam\@accession	MS:1001968	PTM localization score	That term can be there but it is not mandatory	yes
	Corresponding spectrum locus	Reference to experimental spectrum	DataCollection/AnalysisData/SpectrumIdentificationList /SpectrumIdentificationResult/@spectraData_ref			Mandatory attribute	
	Charge state assumed for identification and a measurement of peptide mass error	Assumed charge state	DataCollection/AnalysisData/SpectrumIdentificationList /SpectrumIdentificationResult/SpectrumIdentificationItem /@chargeState			Mandatory attribute	
		Measurement of peptide mass error	DataCollection/AnalysisData/SpectrumIdentificationList /SpectrumIdentificationResult/SpectrumIdentificationItem /@calculatedMassToCharge			This attribute must be present.	
			DataCollection/AnalysisData/SpectrumIdentificationList /SpectrumIdentificationResult/SpectrumIdentificationItem /@experimentalMassToCharge			Mandatory attribute	
	Additional information used for evaluation of confidence		DataCollection/AnalysisData/SpectrumIdentificationList /SpectrumIdentificationResult/SpectrumIdentificationItem /cvParam/@accession			Cannot be validated	

4.Interpretation and validation							
	Methods used for post-processing, re-scoring or re-ranking search engine results.					Cannot be validated	
	Global threshold(s) (or other method) used to accept or reject peptide or protein identifications	For peptides	AnalysisProtocolCollection/SpectrumIdentificationProtocol /Threshold/cvParam/@accession			No additional rules to add.	
		For proteins	AnalysisProtocolCollection/ProteinDetectionProtocol /Threshold/cvParam/@accession			No additional rules to add	
	Results of statistical analysis (if performed)					Cannot be validated	

Tabla 10. Correspondencia entre la información requerida por las directrices MIAPE MSI y los elementos del formato estándar mzIdentML. Columnas: (1) sección MIAPE; (2) información requerida en esa sección; (3) explicación del concepto requerido; (4) ruta XPath al elemento del mzIdentML donde debe estar anotada dicha información; (5) términos de la ontología que deben utilizarse para anotar esa información; (6) nombre de el/los término(s) de la anterior columna; (7) comentarios acerca de cómo validar la información; (8) se muestra si se permite el uso de los términos "hijos" del término de la quinta columna.